Universidade de Lisboa
Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Abordagem Bayesiana para modelar dados com excesso de zeros - aplicação à Parasitologia

João Filipe Azevedo dos Santos

Dissertação
Mestrado em Estatística

2013

Universidade de Lisboa
Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Abordagem Bayesiana para modelar dados com excesso de zeros - aplicação à Parasitologia

João Filipe Azevedo dos Santos

Dissertação
Mestrado em Estatística

Orientadores: Professora Doutora Patrícia de Zea Bermudez (DEIO/FCUL)
e Professora Doutora Luzia Gonçalves (IHMT/UNL)

2013

.

Resumo

Este trabalho pretende, através da abordagem Bayesiana, modelar a carga parasitária de uma doença conhecida como *schistosomose*. Para tal foi analisada a contagem do número de ovos do parasita em amostras de urina de um conjunto de indivíduos proveniente de zonas onde este parasita é endémico.
A *schistosomose* é uma doença cujo hospedeiro definitivo é o homem, sendo responsável por lesões no aparelho urogenital. Estima-se em cerca de 120 milhões de pessoas infectadas (Veles, 2010) e em 600 milhões o número de pessoas em risco de contrair a infecção (Figueiredo, 2008).

Os dados recolhidos são caracterizados por um grande número de indivíduos sem ovos observados na amostra de urina. Quando este número é elevado diz-se que estamos a observar uma variável com "excesso de zeros".

Os Modelos com Excesso de Zeros e os Modelos de duas Partes são usualmente utilizados para lidar com esta situação, dado que têm estruturas específicas que permitem modelar o aparecimento excessivo de zeros face a outros modelos convencionais, usados para modelar variáveis de contagem, que usualmente não possuem estas características.

Os modelos foram implementados nos programas OpenBugs e R, via metodologia MCMC. Foi também efectuada uma comparação com o trabalho já realizado numa perspectiva clássica por Olivença (2011).

Abstract

This work intends to model, using a Bayesian approach, the intensity of a parasitary disease known as schistosomiasis. To do this it was analyzed the number of eggs of the parasite in urine samples from a set of individuals from areas where this parasite is endemic.

Schistosomiasis is a disease whose definitive host is the man, being considered responsible by lesions of the urogenital system. It is estimated that about 120 million persons are infected (Veles, 2010) and 600 millions the number of persons at risk of contracting the infection (Figueiredo, 2008).

The data collected is characterized by a large number of individuals without observed parasite eggs in their urine sample. When this number is high it is said that we are observing a variable with "excess of zeros".

Models with Excess Zeros and Two Part Models are usually used to deal with this characteristics, as they have specific structures that allow modeling the appearance of excessive zeros compared to other conventional models used to model count variables, which usually do not possess these kind of feature. The models were implemented using the programs OpenBugs and R, through MCMC methodology. It was also carried out a comparison with the work already done in a classical perspective by Olivença (2011).

Agradecimentos

Agradeço aos meus pais o seu esforço por me providenciarem uma educação que transformou os meus horizontes e me deu oportunidades que eu não imaginava na minha adolescência.

Queria agradecer em especial, às professoras Patrícia Bermudez e Luzia Gonçalves, pelo seu tempo e interesse ao longo da preparação desta dissertação.

Índice

Lista de Figuras **vii**

Lista de Tabelas **xi**

1 Introdução - Motivação **1**

1.1 Descrição 1

1.1.1 Zeros e Dispersão 3

1.1.2 Abordagem Bayesiana 3

2 A Schistosomose **5**

2.1 Contexto da Doença 5

2.2 Infecção e Ciclo de Vida 6

2.3 Tratamento, Controlo e Prevenção 8

3 Estatística Bayesiana **11**

3.1 Modelos Bayesianos 11

3.1.1 Teorema de Bayes 12

3.1.2 *A Priori* 14

3.1.2.1 Elicitação da distribuição a *priori* 15

3.1.2.2 *Prioris* Conjugadas - *Prioris* Não Informativas 16

3.1.3 *A distribuição Posteriori* 17

3.1.4 Inferência 18

3.1.5 Métodos *MCMC* 18

3.1.5.1 Método de Rejeição 19

3.1.5.2 Algoritmo Metropolis-Hastings 20

3.1.5.3 Método de Amostragem de Gibbs 21

3.1.5.4 Convergência e Diagnóstico das Cadeias de Markov . . 22

4 Definição de Modelos **27**

4.1 Modelos 27

4.1.1 Poisson para dados de Contagens 27

4.1.1.1 Modelo de Regressão Poisson 28

4.1.1.2 Modelo de Regressão Poisson *Zero Inflated* (ZIP) . . . 28

4.1.1.3 Modelo de Regressão Poisson *Zero Altered* (ZAP) ou de *Duas Partes* 30

4.1.2 Binomial Negativa para dados de Contagens 32

4.1.2.1 Modelo de Regressão Binomial Negativo 32

4.1.2.2 Modelo de Regressão Binomial Negativa *Zero Inflated* (ZIBN) 33

4.1.2.3 Modelo de Regressão Binomial Negativa *Zero Altered* (ZABN) ou Modelo *Hurdle* 34

4.1.3 Log-Verosimilhança dos Modelos 35

5 Dados e Covariáveis **37**

5.1 Informação Descritiva 39

5.1.1 Covariáveis 40

6 Resultados dos Modelos **55**

6.1 Definição das distribuições *a Priori* 55

6.2 Resultados 56

6.2.1 Ordenadas Preditivas Condicionais 62

6.2.2 Análise de Resíduos 69

6.2.2.1 Distribuição dos Resíduos Padronizados 70

6.2.2.2 Comparação dos Resíduos com os Valores Esperados . . 71

6.2.3 Binomial Negativa Sobre Parametrizada 75

6.2.3.1 CPO Binomial Negativa Sobre Parametrizada 77

6.2.3.2 Resíduos Binomial Negativa Sobre Parametrizada . . . 78

7 Discussão **81**

Bibliografia **85**

8 Anexos **89**

8.1 Comparação Abordagem Clássica e Bayesiana 89

8.2 Estatísticas das *Posterioris* dos Modelos 91

8.2.1 Estatísticas Poisson GLM 91

8.2.2 Estatísticas Binomial Negativa GLM 92

8.2.3 Estatísticas Modelo ZIP 93

8.2.4 Estatísticas Modelo ZIBN 94

8.2.5 Estatísticas Modelo ZAP 95

8.2.6 Estatísticas Modelo ZANB 96

8.2.7 Estatísticas Modelo Binomial Negativa Sobre Parametrizada . . 97

8.3 Exemplos da Análise de Convergência das Cadeias 98

8.3.1 ZI Binomial Negativa 98

8.4 Definição de uma distribuição arbitrária em Bugs - *Zeros-Ones Trick* . . 104

8.5 Exemplos Programas R - Bugs 105

8.5.1 Programa ZIBN 105

8.5.2 Programa Residuos ZIBN 118

8.6 Programas Auxiliares 126

8.6.1 Programa *HDIofMCMC* 126

8.6.2 Programa *plotPost* 126

8.6.3 Programa *plotChains* 127

Lista de Figuras

2.1 Distribuição da prevalência da *schistosomose* 5
2.2 Ciclo de vida do *schistosoma* 7

3.1 Ilustração da partição do universo 13
3.2 Ilustração do Teorema das Probabilidades Totais 14
3.3 Ilustração do Método de Rejeição 19
3.4 Exemplo do traço de três cadeias com comportamento aleatório 23
3.5 Exemplo do traço de três cadeias em que se verifica um comportamento não aleatório 23
3.6 Gráfico Brooks-Gelman-Rubin (BRG) - exemplo de convergência de $\hat{R}$ para 1 25
3.7 Gráfico Brooks-Gelman-Rubin (BRG) - exemplo de falha de convergência de $\hat{R}$ para 1 25
3.8 Exemplo de elevada autocorrelação - *thinning 0* 26
3.9 Exemplo de baixa autocorrelação - *thinning 35* 26

4.1 Exemplos da f.m.p. do modelo ZIP 30
4.2 Exemplos da f.m.p. do modelo ZAP 31
4.3 Exemplos da f.m.p. do modelo ZINB 33
4.4 Exemplos da f.m.p. do modelo ZANB 34

5.1 Mapa de Angola 38
5.2 Diagrama de barras do número de ovos observado por amostra de urina 39
5.3 Número de ovos observado em função da Idade 41
5.4 *Box-Plots* do número de ovos observado segundo o Género 42
5.5 *Box-Plots* do número de ovos observado segundo a Residência 43

LISTA DE FIGURAS

5.6 *Box-Plots* do número de ovos observado segundo a Naturalidade 44
5.7 *Box-Plots* do número de ovos observado segundo a Profissão 45
5.8 Distribuição dos indivíduos por Género e Profissão 46
5.9 *Box-Plots* do número de ovos observado segundo o Conhecimento da Doença 47
5.10 Distribuição do nível de Conhecimento da Doença por Profissão 48
5.11 *Box-Plots* do número de ovos observado segundo o resultado do Teste Hematúria (ou de Hematúria) 49
5.12 *Box-Plots* do número de ovos observado segundo o Local de Contacto com a Água 50
5.13 *Box-Plots* do número de ovos observado segundo o Motivo de Contacto com a Água 51
5.14 Distribuição dos indivíduos por Local de Contacto segundo o Motivo de Contacto com Água 52
5.15 *Box-Plots* do número de ovos observado segundo a Existência de Água Canalizada 53
5.16 *Box-Plots* do número de ovos observado segundo a Existência de WC dentro de Casa 54

6.1 *Posteriori* da Variável Género: Masculino - Poisson GLM 59
6.2 *Posteriori* da Variável Género: Masculino - ZIP 59
6.3 *Posteriori* da Variável Género: Masculino - ZAP 59
6.4 *Posteriori* da Variável Idade - Poisson GLM 59
6.5 *Posteriori* da Variável Idade - Binomial Negativa GLM 59
6.6 *Posteriori* da Variável Hematúria: Teste Positivo - ZIBN - Médias . . . 60
6.7 *Posteriori* da Variável Hematúria: Teste Positivo - ZIBN - Zeros 60
6.8 *Posteriori* da Variável Tem Conhecimento da Doença: Não Tem - Binomial Negativa GLM 61
6.9 *Posteriori* da Variável Tem Conhecimento da Doença: Não Tem - ZIBN 61
6.10 *Posteriori* da Variável Tem Conhecimento da Doença: Não Tem - ZABN 61
6.11 *Posteriori* da Ordenada na Origem - ZIBN - Zeros 61
6.12 *Posteriori* da Variável Género: Masculino - ZIBN - Zeros 61
6.13 *Posteriori* da Variável Idade - ZIBN - Zeros 61

6.14 CPO Poisson por número de ovos observado por indivíduo 63
6.15 CPO ZIP por número de ovos observado por indivíduo 63
6.16 CPO ZAP por número de ovos observado por indivíduo 64
6.17 CPO Binomial Negativa GLM por número de ovos observado por indivíduo 64
6.18 CPO ZIBN por número de ovos observado por indivíduo 64
6.19 CPO ZABN por número de ovos observado por indivíduo 65
6.20 Log(CPO) Poisson GLM por Indivíduo 66
6.21 Log(CPO) Binomial Negativa GLM por Indivíduo 66
6.22 Log(CPO) ZIP por Indivíduo 66
6.23 Log(CPO) ZIBN por Indivíduo 66
6.24 Log(CPO) ZAP por Indivíduo 67
6.25 Log(CPO) ZABN por Indivíduo 67
6.26 Histograma dos Resíduos Padronizados 70
6.27 Resíduos Padronizados Poisson GLM *contra* Valores Esperados 71
6.28 Resíduos Padronizados Binomial Negativa GLM *contra* Valores Esperados 72
6.29 Análise Resíduos Padronizados ZIP *contra* Valores Esperados 72
6.30 Análise Resíduos Padronizados ZIBN *contra* Valores Esperados 73
6.31 Análise Resíduos Padronizados ZAP *contra* Valores Esperados 73
6.32 Análise Resíduos Padronizados ZABN *contra* Valores Esperados 74
6.33 CPO Binomial Negativa Sobre Parametrizada por Número de ovos observado 77
6.34 Log(CPO) Binomial Negativa Sobre Parametrizada por Indivíduo 77
6.35 Histograma dos resíduos Binomial Negativa Sobre Parametrizada 78
6.36 Resíduos Binomial Negativa Sobre Parametrizada *contra* Valores Esperados 79
6.37 Resíduos Binomial Negativa Sobre Parametrizada por Logaritmo *contra* Valores Esperados 79

LISTA DE FIGURAS

Lista de Tabelas

4.1 Resumo dos modelos utilizados 27
4.2 Log-Verosimilhanças dos Modelos 35

5.1 Agregação da Zona de Naturalidade 38
5.2 Estatísticas do número de ovos observado por amostra de urina 39
5.3 Estatísticas dos indivíduos por Idade 40
5.4 Estatísticas segundo o Género dos indivíduos 42
5.5 Estatísticas segundo a Província de Residência 43
5.6 Estatísticas segundo a Naturalidade 44
5.7 Estatísticas segundo a Profissão 45
5.8 Idade média dos indivíduos por Profissão 46
5.9 Distribuição dos indivíduos por Género e Profissão 46
5.10 Estatísticas segundo o Conhecimento da Doença 47
5.11 Distribuição dos indivíduos em função do Conhecimento da Doença por Profissão 47
5.12 Análise do número médio de ovos dos indivíduos infectados por Profissão em função do Conhecimento da Doença 48
5.13 Estatísticas segundo os resultados do Teste Hematúria 49
5.14 Estatísticas segundo o Local de Contacto com a Água 50
5.15 Estatísticas segundo o Motivo de Contacto com a Água 51
5.16 Distribuição dos indivíduos por Local de Contacto segundo o Motivo de Contacto com Água 52
5.17 Estatísticas segundo a Existência de Água Canalizada 53
5.18 Estatísticas segundo a Existência de WC Dentro de Casa 54

LISTA DE TABELAS

6.1 Resumo dos Modelos 57
6.2 Médias a *posteriori* dos parâmetros dos modelos 58
6.3 Distribuição por categoria de valores do logaritmo das CPO 68
6.4 Distribuição por categoria de valores do logaritmo das CPO dos Indivíduos sem ovos do parasita observados 68
6.5 *Negative cross-validatory Log Likelihood* (NLL) 69
6.6 Análise Estatística dos Resíduos 70
6.7 Valor médio da *posteriori* dos parâmetros do modelo Binomial Negativo Sobre Parametrizado 76
6.8 Distribuição por categoria de valores do logaritmo das CPO do modelo Binomial Negativo Sobre Parametrizado e Estimativa NLL 77
6.9 Estatística descritiva dos Resíduos Modelo Binomial Negativo Sobre Parametrizado 78

8.1 Valores dos parâmetros dos modelos Clássico *contra* Bayesiano 90
8.2 Estatísticas Poisson GLM 91
8.3 Estatísticas Binomial Negativa GLM 92
8.4 Estatísticas ZIP 93
8.5 Estatísticas ZIBN 94
8.6 Estatísticas ZAP 95
8.7 Estatísticas ZABN 96
8.8 Estatísticas Binomial Negativa Sobre Parametrizada 97

1

Introdução - Motivação

1.1 Descrição

A *schistosomose* é uma doença parasitária em que um dos hospedeiros é o homem. Está associada a lesões do aparelho urogenital, alterações do trato intestinal, anemia, falência renal e inclusive à presença de cancro da bexiga (Shiff, 2006 e Santos et. al., 2012).

Os sintomas desta doença são causados principalmente pela deposição de ovos do parasita em tecidos nas proximidade da bexiga e intestinos e pela respectiva resposta imunitária (Mahmoud, 2001).

Este trabalho centra-se na análise da espécie *schistosoma haematobium* (*s. haematobium*), cujos ovos são expelidos através da urina, sendo esta espécie responsável pela *schistosomose* urinária que provoca lesões no aparelho urogenital (Figueiredo, 2008). Estima-se em cerca de 120 milhões de pessoas infectadas pelo *s. haematobium* (Veles, 2010) em 600 milhões em risco de contrair a infecção (Figueiredo, 2008).

Dado o nível de exposição ao parasita e à morbidade da população infectada é importante conhecer os mecanismos de infecção e a situação dos infectados de forma a poder desenvolver medidas de prevenção, diagnóstico e tratamento.

Seguindo uma abordagem *Bayesiana*, pretende-se modelar a distribuição da carga parasitária associada ao parasita *s. haematobium* de uma amostra da população Angolana das províncias de Bengo, Luanda e Kwanza Sul, recolhida e cedida pela Dr. Jacinta Teresa no âmbito da sua dissertação de mestrado em Parasitologia Médica do

IHMT (Figueiredo, 2008). Será também efectuada uma comparação com a análise já desenvolvida por Olivença (2011) numa perspectiva *Clássica*.

A carga parasitária é avaliada de forma aproximada pela contagem do número de ovos do parasita existente nas amostras de 10 ml de urina recolhidas (Figueiredo, 2008). Nas amostras destes indivíduos verificou-se existir uma grande quantidade nas quais não se encontram ovos (*zeros*), uma frequência elevada de contagens pequenas e indivíduos com uma contagem particularmente elevada de ovos (Figueiredo, 2008 e Cardoso, 2010).

À modelação de variáveis de *contagem* é usualmente associada a distribuição de *Poisson*. No entanto, esta não consegue resolver as dificuldades levantadas por um número elevado de *zeros* e lidar com a *sobredispersão*, características da distribuição do número de ovos nas amostras de urina em análise neste trabalho.
Sem a devida modelação desta variabilidade extra, a precisão dos parâmetros será sobre-estimada e os intervalos de credibilidade serão demasiado pequenos (Cameron & Trivedi, 1998, citado em Condom, 2006).

De modo a dar resposta a estas características, são usados os modelos com Excesso de Zeros *(Zero Inflated Models (ZI)) (Lambert, 1992; Greene, 1994)* e os modelos de duas Partes *(Zero Altered (ZA) ou Hurdle Models) (Mullahy, 1986; Heilbron, 1989)*.
Estes permitem transformar os modelos usuais, de forma a capacitá-los para modelar variáveis de "contagem" que apresentam um número excessivo de *zeros*.
Também será utilizada a distribuição Binomial Negativa como alternativa à Poisson com o objectivo de obter modelos menos restritos no que respeita à dispersão (Zuur, 2009)

1.1.1 Zeros e Dispersão

É possível estabelecer à partida algumas razões para a existência de *excesso de zeros* na contagem de ovos na amostra de urina. Podem ser parte integrante do fenómeno que está a ser analisado, a urina ter sido recolhida num período de incubação do parasita não sendo ainda possível detectar a infecção (Cardoso, 2010) ou simplesmente erros de contagem (Zuur, 2009).

Outra particularidade importante na distribuição da contagem de ovos é a *sobresdispersão*, que pode resultar da existência de valores extremos, da agregação de grupos de indivíduos com valores altos e baixos ou problemas de amostragem (não aleatoriadade) (Efron, 1986 citado em Condom, 2006).

Neste trabalho propomo-nos usar modelos semelhantes ou transformações de modelos *clássicos*, como os modelos ZI e ZA, de forma a modelar o número excessivo de *zeros*. Também é utilizada a *Binomial Negativa* para modelar variáveis de *contagem* em alternativa à distribuição *Poisson*, por ser mais flexível no que respeita à capacidade de representar a variabilidade presente nos dados.

1.1.2 Abordagem Bayesiana

Neste trabalho considera-se um conjunto de dados já analisado por Olivença (2011) numa perspectiva frequencista, sendo agora proposta uma abordagem Bayesiana. Assim serão explorados alguns modelos com recurso à metodologia *Monte Carlo Markov Chain* (*MCMC*) e a software gratuito (R e OpenBugs).

2

A Schistosomose

2.1 Contexto da Doença

A *schistosomose* é a segunda doença parasitária humana mais estudada, a seguir à malária (Mahmud, 2001). É reconhecida pela World Health Organization (WHO) como sendo uma doença tropical negligenciada (NTD). A sua permanência deve-se em parte à pobreza e falta de medidas preventivas necessárias por parte de entidade responsáveis nos países onde esta doença persiste (Bruun et al., 2008).

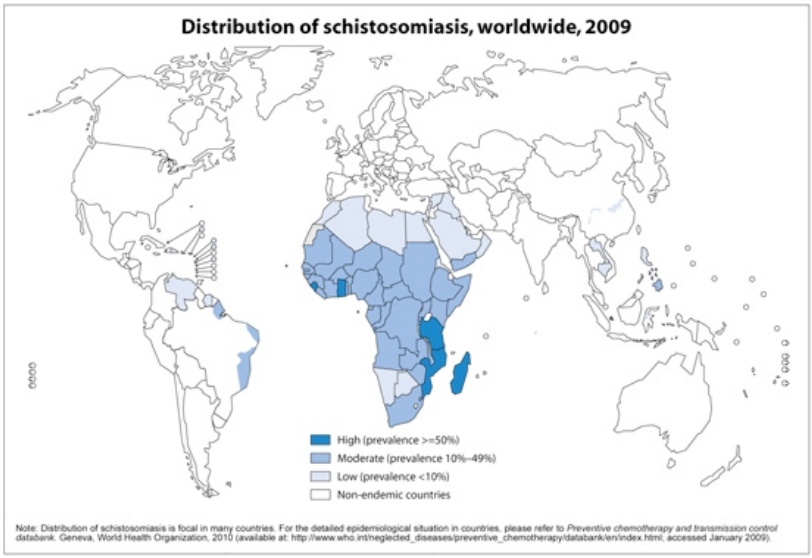


Figura 2.1: Distribuição da prevalência da *schistosomose*.
Retirado e adaptado de *http://www.neglecteddiseases.gov/target_diseases/schistosomiasis* : Dados e Mapa obtidos de *WHO Map Library*

A *schistosomose* é endémica em mais de 70 países e territórios distribuídos por África, América do Sul e Ásia (Chitsulo et al., 2000). É uma doença encontrada em países afectados por pobreza extrema, em áreas com pouco ou nenhum acesso a água tratada e sem cuidados sanitários adequados, sendo particularmente prevalente em África sub-Sahariana e em certas áreas rurais da China (Bruun et al. 2008). Um relatório epidemiológico recente da WHO indica que existem mais de 240 milhões de pessoas que requerem tratamento (WER, 2013). Esta doença está associada a lesões do aparelho urogenital, trato intestinal, anemia, falência renal e inclusive ao cancro da bexiga (Shiff, 2006 e Santos et. al. 2012). Dado o número elevado de pessoas infectadas ou expostas à infecção e à morbidade da doença, é considerada um grave problema de saúde pública, em particular em África (WER, 2013).

2.2 Infecção e Ciclo de Vida

Existem várias espécies do parasita *schistosoma*: *s. haematobium*, *s. intercalatum*, *s. japonicum*, *s. mansoni* e *s. mekongi*. Apesar de serem algo diferentes, as espécies com relevância médica (*s. haematobium*, *s. japonicum* e *s. mansoni*) têm ciclos de vida semelhantes (Mahamud, 2011). O ciclo de vida deste parasita está dividido em duas fases: uma no organismo humano (o hospedeiro definitivo) e outra no interior de um molusco de água doce (o hospedeiro intermediário) (Santos et. al. 2012).

No caso do *s. haematobium*, o hospedeiro intermédio é o caracol da espécie *Bulinus*. Estes moluscos vivem em pequenas porções de águas paradas ou de fraca corrente ou mesmo em reservatórios artificiais de água. Quando infectados, os moluscos libertam cercárias (formas larvais do parasita) que nadam na água circundante expondo as pessoas à infecção ao entrarem em contacto com os parasitas nestes locais. Por esta razão a *schistosomose* é também chamada de "febre dos caracóis" (Mahmud, 2001).

Os ovos de *schistosoma* libertados nas excreções humanas, quando entram em contacto com água doce e em condições adequadas (de temperatura, luminosidade e pressão osmótica), eclodem, libertando o *miracídio* (Cardoso, 2010), uma forma larvar que possui cílios de forma a movimentar-se na água. Os *miracídios* continuam a sua evolução após penetrarem/infectarem os moluscos de água doce. Dentro dos moluscos, o *miracídio* desenvolve-se para uma forma sacular denominada *esporocisto*, que se reproduz

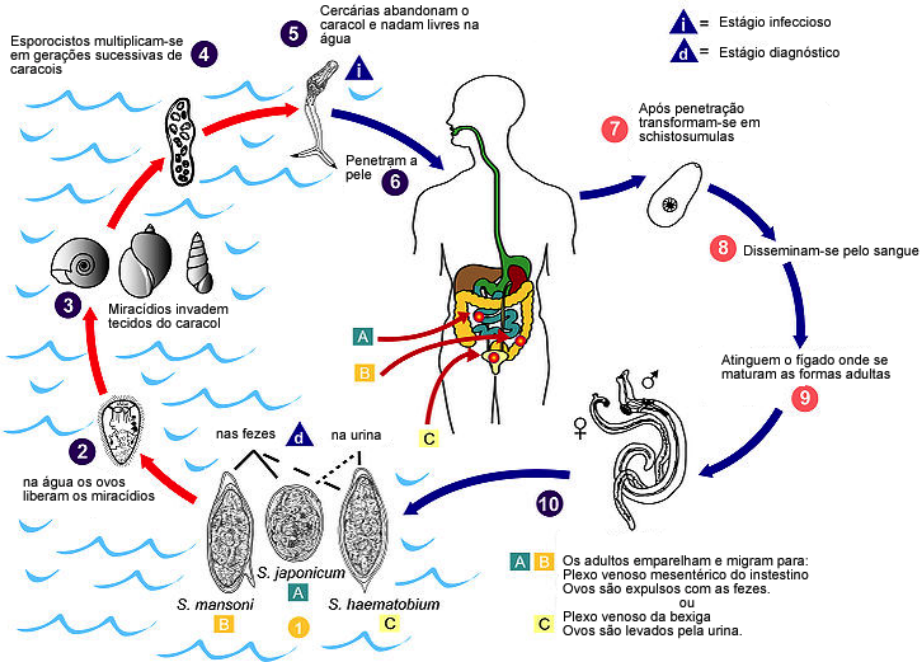


Figura 2.2: Ciclo de vida do *schistosoma*.
Retirado e adaptado de *http://pt.wikipedia.org/wiki/Schistosoma*

de forma assexuada dando origem às cercárias. Um *miracídio* pode dar origem a milhares de cercárias (Cardoso, 2010).

As cercárias são um novo estágio larvar do parasita. Estas abandonam o molusco, ficando dispersas nas águas circundantes. Ao penetrarem na pele (nua e sem feridas) ou mucosas das pessoas, evoluem para um novo estágio denominado *schistosómulos*, sendo que estes organismos alcançam a corrente sanguínea, passando pelos pulmões, coração até chegar ao fígado. No fígado maturam para as suas formas adultas, dando-se o acasalamento entre machos e fêmeas. Posteriormente, o par de vermes (macho e fêmea) desloca-se para os locais onde se dará a libertação dos ovos embrionados. No caso do *s. haematobium*, o local principal onde se dará a libertação dos ovos será o plexo venoso vesical, em particular em torno dos ureteres e da bexiga (Mahmoud, 2001).

Dos ovos libertados do *s. haematobium*, há os que atravessam as paredes dos ureteres ou bexiga e são posteriormente expelidos através da urina recomeçando o ciclo, enquanto outros são retidos dentro do corpo das pessoas infectadas. A deposição destes ovos nos tecidos como a bexiga, ureteres e rins e a respectiva resposta imunitária

(irritação, inflamação, hematúria) dá origem à patologia desta doença parasitária. A irritação e inflamação pode levar ao desenvolvimento de lesões crónicas e agudas do tracto urinário (Mahmoud, 2001). A *schistosomose urinária* é mesmo considerada como uma das causas principais de hematúria no mundo (Mahamud, 2001).

As pessoas infectadas excretam ovos em (ou perto de) colecções de água como lagos, rios e canais, expondo à infecção agricultores, pescadores, crianças que brincam e mesmo pessoas em tarefas domésticas como lavar roupa. As crianças também devido a hábitos de higiene inadequados, tornam-se assim particularmente susceptíveis à infecção (Cardoso, 2010).

Verifica-se que a infecção e a sua prevalência é determinada por hábitos de higiene dos indivíduos, acesso a estruturas sanitárias e reservas de água "limpa", à existência dos moluscos nas colecções de água próximas das populações e à resistência dos indivíduos à infecção.

2.3 Tratamento, Controlo e Prevenção

O tratamento médico dos infectados é feito principalmente através de quimioterapia. Existem vários medicamentos, sendo utilizado principalmente um medicamento administrado oralmente chamado Praziquantel, que é indicado para os vários tipos de *schistosoma* (Cardoso, 2010). A sua utilização leva à redução da carga parasitária e de outros indicadores da doença como a hematúria. A administração em massa destes fármacos não apresentam solução definitiva visto que a taxa de reinfecção é elevada e não resolve a questão da transmissão da doença (McManus & Loukas, 2008). Uma breve exposição a água infestada com cercárias é suficiente para efectivar a transmissão da doença e um só indivíduo infectado pode contaminar uma colecção de água e transformar essa zona num local de elevada probabilidade de transmissão durante vários meses (King, 2010).

É relevante referir que sem o hospedeiro intermédio a disseminação da doença e a infecção não seriam possíveis, de modo que o desenvolvimento de medidas de controlo dos moluscos que servem de hospedeiros intermédios também constituiem um modo de

prevenção.

Existem estudos sobre a imunoresistência dos indivíduos à infecção de forma a possibilitar o desenvolvimento de uma vacina. Estes estudos suportam a possibilidade de implementar esta solução. A vacinação é considerada um passo importante visto que, apesar da elevada eficácia dos fármacos utilizados, os indivíduos infectados continuam a excretar ovos que reiniciam um novo ciclo parasitário expondo a população à reinfecção, havendo também a possibilidade de se desenvolver resistência aos medicamentos actualmente utilizados (McManus & Loukas, 2008).

3

Estatística Bayesiana

Neste capítulo pretende-se fazer uma curta descrição sobre a Estatística Bayesiana e metodologia *MCMC* como ferramenta para os desenvolvimento de modelos Bayesianos.

3.1 Modelos Bayesianos

Seja Y uma variável aleatória com função densidade de probabilidade (f.d.p) $p(y \mid \theta)$ onde θ é o respectivo vector de parâmetros. Tendo em conta a informação observada, $y_1, y_2, ..., y_n$, um conjunto de observações de n variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), $Y_1, Y_2, .., Y_n$, com f.d.p. $p(y \mid \theta)$, o modelo conjunto ou distribuição conjunta

$$L(\theta \mid y) = \prod_{i=1}^{n} p(y_i \mid \theta)$$

é denominada por verosimilhança, também representada equivalentemente por $p(y \mid \theta)$ (Congdon, 2003). Na *Estatística Clássica* é usualmente utilizada a verosimilhança que reúne a informação dos dados observados e da estrutura probabilística do modelo de forma a poder fazer inferências sobre θ. A abordagem *Bayesiana* introduz mais uma componente, a informação *a priori* acerca dos parâmetros θ. Nesta perspectiva, a informação *a priori* representa o conjunto de evidências ou suposições existentes acerca dos parâmetros do modelo antes de observar os dados em análise. θ é considerado como um vector de variáveis aleatórias com distribuição de probabilidade $\pi(\theta)$, a qual reflecte a nossa incerteza acerca dos parâmetros do modelo.

Se bem que por uma ordem lógica foi descrito que *é adicionada uma componente* de

modo a fazer comparação entre as duas abordagens, *Clássica e Bayesiana*, entenda-se que a informação $p(y \mid \theta)$ é usada para actualizar o nosso conhecimento da informação *a priori* acerca dos parâmetros θ (Congdon, 2003). O resultado da combinação entre a informação dos dados actuais e a informação *a priori* é denominada *posteriori*, sendo representada por $\pi(\theta \mid y)$. A *posteriori* é um elemento muito importante na qual se baseia toda a inferência Bayesiana (Paulino et. al., 2003).

3.1.1 Teorema de Bayes

A *Estatística Bayesiana* está fundamentada numa interpretação do *Teorema de Bayes*. Este permite dentro da sua estrutura probabilística realizar a actualização da informação que possuímos presentemente após a observação de dados adicionais. Nesta subsecção pretende-se abordar de forma resumida como o *Teorema de Bayes* dá suporte à metodologia *Bayesiana*, que é um misto de conhecimento *a priori*/subjectivo e da informação adicional disponível fornecida pelos dados.

Sejam A e B são dois acontecimentos possíveis do universo Ω. Seja P(B) a probabilidade de ocorrência de B e seja $\bar{B}$ o acontecimento complementar de B donde P($\bar{B}$) = 1- P(B).
Segundo a definição de probabilidade condicional ou *Regra de Bayes* temos:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

Pela regra da multiplicação obtém-se:

$$P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B). \tag{3.1}$$

Verifica-se que $A = (A \cap B) \cup (A \cap \bar{B})$ onde $(B \cup \bar{B}) = \Omega$ e portanto

$$P(B \mid A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap \bar{B})}$$

se recorrermos a (3.1), substituindo verifica-se que

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})}. \tag{3.2}$$

O acontecimento A é obtido através da *união* de dois acontecimentos disjuntos $(A \cap B)$ *e* $(A \cap \bar{B})$, onde a união de B e $\bar{B}$ é todo universo considerado, pelo que estes definem uma partição de Ω. Assuma-se agora um conjunto de acontecimentos $B_{i,\ i=1,..,n}$ e que estes definam uma partição de Ω ou seja:

- $B_1 \cup B_2 \cup ... \cup B_n = \Omega$
- $B_i \cap B_j = \emptyset$ *para qualquer* $i \neq j$ *onde* $i, j \in \{1, 2, ...n\}$

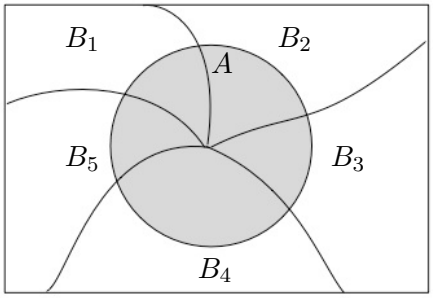


Figura 3.1: Ilustração da partição do universo

Assim pode-se generalizar que $A = (A \cap B_1) \cup (A \cap B_2) \cup ... \cup (A \cap B_n)$. Se recorrermos a (3.2) temos

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{\sum_{j=1}^{n} P(A \cap B_j)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}. \tag{3.3}$$

O denominador da expressão em (3.3) é denominado Teorema das Probabilidade Totais. A soma das probabilidades de ocorrer A dado $B_{i,\ i=1,2,...,n}$ ponderado pela probabilidade ou credibilidade P(B_i) dá-nos a probabilidade do acontecimento A.

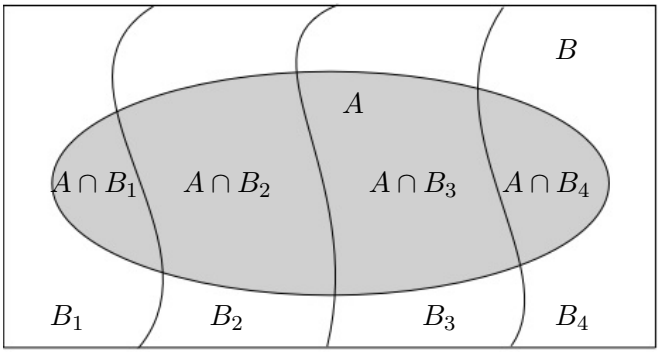


Figura 3.2: Ilustração do Teorema das Probabilidades Totais

Tendo em conta os seguintes pressupostos, se $P(B_i)_{\ i=1,2,\ldots,n}$ representar um conjunto de informações *a priori*, subjectivas sobre os acontecimentos não observáveis $B_{i\ i=1,2,\ldots,n}$ e se A for um acontecimento adicional observado, então o resultado final poderá ser interpretado da seguinte forma: $P(B_i \mid A)$ será a actualização das suposições sobre B_i dada a informação obtida A (William, 2004 e Paulino et. al., 2003).

$P(B_i \mid A)$ contém o "peso" de cada B_i que representará informação subjectiva, mas também é baseada na "nova" informação observada A.

3.1.2 *A Priori*

Na análise *Estatística Clássica* os parâmetros θ são tomados como características fixas dos modelos. Em contrapartida na *abordagem Bayesiana* é assumida incerteza acerca destes. Como foi já referido, os parâmetros são considerados variáveis aleatórias de forma a representar o grau de credibilidade que é atribuído a θ (Paulino et. al., 2003).

A informação *a priori* é então formalizada como uma variável aleatória θ com função de distribuição $\pi(\theta)$, denominada por *distribuição a priori* ou simplesmente *priori*. Esta serve de suporte para modelar os pressupostos acerca dos parâmetros θ. A *priori* é definida inicialmente no processo de construção dos modelos antes de considerar os dados em análise. A especificação da *priori* é bastante importante pois tem "peso" nos resultados dos modelos.

3.1.2.1 Elicitação da distribuição a *priori*

Elicitação da *priori* é a formalização da estrutura funcional ou paramétrica da *priori* $\pi(\theta)$. Uma situação que ilustra este "trabalho" será a discussão entre o Estatístico e um especialista, em que o primeiro questiona acerca da possível estrutura da informação e do nível de credibilidade que o especialista tem em determinadas ocorrências (Paulino et. al. 2003). Esta informação também poderá ser obtida através de outros estudos relacionados, tais como meta análise formal ou informal (Congdon, 2003).
Alguns exemplos de técnicas usadas para formalizar este conhecimento são:

- O Método do Histograma - Para um conjunto de valores possíveis de θ, constrói-se uma representação gráfica da distribuição de $\pi(\theta)$. Esta representação será depois avaliada de forma a estar de acordo com o conhecimento que se pretende formalizar e de forma a ajustar uma distribuição probabilística a este histograma.

- Elicitação dos Hiperparâmetros - Quando existe previamente ou está bem estabelecida uma estrutura funcional de $\pi(\theta)$, bastará discutir com o especialista uma parametrização conveniente. No entanto, o especialista poderá não ter conhecimentos acerca de cálculo de probabilidades para dar um bom *input* pelo que se poderá, por exemplo, analisar *quantis* de probabilidade de forma a que o analista consiga obter possível estrutura de parâmetros.

- No Método Preditivo de elicitação questiona-se um especialista sobre valores observados y^* mas cujos dados não estão disponíveis para um modelo $p(y \mid \theta)$ já definido pelo analista. Posteriormente procura-se identificar uma *priori* $\pi(\theta)$ que se adapte de forma a obter

$$p(y^*) = \int p(y \mid \theta)\pi(\theta)d\theta.$$

  Nesta situação, a necessidade analítica de inversão para poder definir uma distribuição $\pi(\theta)$ adequada é dificultada pela necessidade de que $p(y \mid \theta)\pi(\theta)$ seja tratável de forma analítica.

Estas e outras metodologias de elicitação de *prioris* poderão ser encontradas em Paulino et. al. (2003).

3.1.2.2 *Prioris* Conjugadas - *Prioris* Não Informativas

A obtenção de uma *posteriori* analiticamente tratável nem sempre é fácil, pelo que nos primeiros modelos Bayesianos eram evitadas *prioris* não tratáveis ou não *conjugadas* (Ntzoufras, 2009).

Seja $\pi(\theta)$ uma *priori* membro de uma familia de distribuições $\mathcal{F}$ com parâmetros α. Esta é conjugada de $p(y \mid \theta)$ se a *posteriori* $\pi(\theta \mid y)$ também pertence a $\mathcal{F}$. Portanto se $\theta \sim \mathcal{F}(\alpha)$ então *a posteriori* $\theta \mid Y \sim \mathcal{F}(\bar{\alpha})$, onde $\bar{\alpha}$ são os parâmetros *a posteriori* de $\mathcal{F}$. O que significa que a facilidade de análise que possuímos da *priori* é mantida quando analisamos a *posteriori* (Ntzoufras, 2009).

Apesar de que a selecção de distribuições *a priori* é usualmente guiada pelo suporte de valores que esta pode ter, a utilização de distribuições *a priori* conjugadas goza de propriedades que permitem uma actualização mais fácil, simplifica o processo de estimação e usualmente são suficiente capazes para descrever a informação *a priori* (Rowe, 2003).

Quando não existe informação *a priori* relevante ou esta é pouco significativa em relação à informação amostral recorre-se a *prioris/distribuições não informativas* (Paulino et. al., 2003) ou *vagas*. Estas pretendem reproduzir o nível de ignorância no que respeita à informação que possuímos à partida do modelo (Paulino et. al., 2003). Neste sentido pretendem-se distribuições com pouco "peso" na *posteriori*.

As distribuições *a priori*, no contexto desta análise, devem ser distribuições cujos suportes contêm o conjunto de valores possível dos parâmetros que estamos a modelar, são definidas com um nível de variância elevado de forma a representar a nossa incerteza e por isso são denominadas *priori pouco informativas* ou *minimamente informativas* (Ntzoufras, 2009). Isto significa que terão muito pouco peso ou um peso negligenciável no resultado final do modelo e a informação dos dados actuais terá um papel mais importante. Neste sentido a *Estatística Bayesiana* aproxima-se da abordagem *Clássica*, com a ressalva que se admite a ignorância sobre os parâmetros do modelo e essa mesma falta de informação é introduzida no modelo (Paulino et. al., 2003).

3.1.3 *A distribuição Posteriori*

O objectivo principal é obter a distribuição dos parâmetros do modelo, o resultado probabilístico é denominado *distribuição a posteriori* ou *posteriori* e é representada usualmente por $\pi(\theta \mid y)$, como já foi mencionado. Esta distribuição contém toda a informação actual sobre os parâmetros θ (Gelman et. al., 2003). Para obter a *posteriori* recorre-se ao *Teorema de Bayes* e ao cálculo probabilístico, sem recurso a resultados assimptóticos.

A densidade conjunta da estrutura dos parâmetros e dos dados por ser obtida tendo em conta que

$$p(y, \theta) = p(y \mid \theta)\pi(\theta) = \pi(\theta \mid y)p(y),$$

assim a distribuição *a posteriori* será

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)},$$

onde o denominador $p(y)$ é obtido utilizando o *Teorema das Probabilidade Totais*. Sendo também denominado como a verosimilhança marginal dos dados (Congdon, 2003),

$$p(y) = \int p(y \mid \theta)\pi(\theta)d\theta. \tag{3.4}$$

Considerando (3.4) como uma constante normalizadora podemos indicar que

$$\pi(\theta \mid y) \propto p(y \mid \theta)\pi(\theta),$$

ou seja, a distribuição *a posteriori* é uma função da informação *a priori* e da evidência que se encontra nos dados (Congdon, 2003). A *posteriori* é uma *síntese* das duas fontes de informação consideradas. Como se verifica, *a priori* $\pi(\theta)$ tem "peso" no *resultado final* pelo que deve ser escolhida ou elicitada cuidadosamente.

3.1.4 Inferência

Depois de obtida a distribuição *posteriori* $\pi(\theta \mid y)$, o objectivo de interesse seguinte será analisar o conjunto de parâmetros θ que a caracteriza. Neste sentido podemos pretender obter a:

- Distribuição Marginal *a posteriori* de um parâmetro, $\pi(\theta_i \mid y) = \int \pi(\theta \mid y)\ d\theta_{-i}$, com $\theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n)$.
- Média *a posteriori* $E(\theta_i \mid y) = \int \theta_i\ \pi(\theta \mid y)\ d\theta_i$.
- Quantis de probabilidade.

Salvo casos específicos, como distribuições conjugadas, o tratamento analítico de $\pi(\theta \mid y)$ é complexo, pelo que se recorre a métodos numéricos ou aproximações (Congdom, 2003 e Paulino et. al., 2003).
Esta tem sido uma das maiores dificuldades da *Estatística Bayesiana*, mas com o desenvolvimento computacional alguns destes entraves têm vindo a ser ultrapassados. É possível recorrer a diferentes métodos de simulação, em particular à metodologia *MCMC*, a qual permite obter amostras das distribuições que pretendemos analisar. Esta metodologia ultrapassa a necessidade de integração para obter uma distribuição marginal ou o valor esperado de um parâmetro, permitindo gerar conjuntos de dados de distribuições multivariadas complexas.

Os métodos *MCMC* utilizam várias técnicas que permitem simular a *distribuição* de todos os parâmetros $\pi(\theta_1, \theta_2, .., \theta_n \mid y)$. Desta forma é possível analisar *a posteriori* dos modelos com recurso a amostras geradas por estes métodos e inferir sobre θ.

3.1.5 Métodos *MCMC*

Pretende-se gerar amostras da distribuição $\pi(\theta \mid y)$ da qual não o podemos fazer directamente. Suponha-se que é possível construir uma cadeia de Markov com espaço de estados Θ em que a sua distribuição de equilíbrio é $\pi(\theta \mid y)$. Se gerarmos estados de uma cadeia nestas condições, o conjunto de valores obtidos, em condições de estacionariedade, podem ser usados como uma amostra que permitirá inferir sobre as características do modelo $\pi(\theta \mid y)$. A dificuldade prende-se em como construir esta cadeia de Markov.

3.1.5.1 Método de Rejeição

Nesta sub-secção é introduzido de forma resumida um método abrangente de simulação de distribuições de forma indirecta e que é utilizado por métodos mais complexos que serão descritos seguidamente (Gelman et. al. 2003). Esta metodologia baseia-se no princípio de que uma amostra de uma distribuição $\pi(\theta)$ pode ser obtida através de amostragem uniforme de valores sob a função de densidade de θ. Para realizar a amostragem é requerida uma função auxiliar ou de referência q(θ), positiva e definida em todo o espaço de θ. Supõe-se também que é possível obter uma amostra aleatória de q(θ) e que o rácio $\frac{\pi(\theta)}{q(\theta)}$ é limitado superiormente por M, sendo M uma constante positiva. O algoritmo para obter a amostra é então definido da seguinte forma (Gelman et. al. 2003):

1. Amostrar $\theta_\star$ de $q(\theta)$
2. Amostrar μ de uma distribuição $U(0,1)$ de forma a que seja independente de $\theta_\star$
3. Se $\mu \cdot Mq(\theta_\star) < \pi(\theta_\star)$ aceitar $\theta_\star$ como valor da distribuição $\pi(\theta)$, caso contrário rejeitar $\theta_\star$

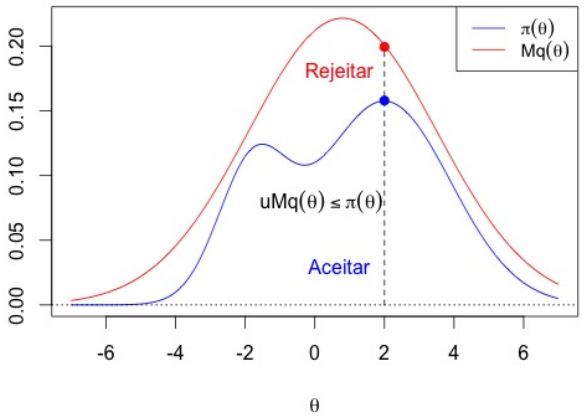


Figura 3.3: Ilustração do Método de Rejeição
Adaptação de gráfico das aulas de Estatística Bayesiana (FCUL DEIO)

Em termos gráficos está a ser criado um par $(\theta_\star, \nu)$ com $\nu = \mu \cdot Mq(\theta_\star)$. O valor de ν está uniformemente distribuído sob $Mq(\theta_\star)$ e são aceites os pontos tais que $\nu < \pi(\theta_\star)$. Prova-se que uma amostra obtida com este método tem distribuição $\pi(\theta)$.

3.1.5.2 Algoritmo Metropolis-Hastings

O algoritmo *Metropolis* permite criar uma cadeia de Markov cuja distribuição de equilíbrio é $\pi(\theta \mid y)$, sem recurso a um método directo de simulação. É uma adaptação de um passeio aleatório que através do método de rejeição converge para uma determinada distribuição $\pi(\theta \mid y)$. O algoritmo pode ser descrito da seguinte forma:

1. Designar um ponto inicial θ_0 de modo que $\pi(\theta_0 \mid y) > 0$.
2. Fazer $\theta_{actual} = \theta_0$.
3. Amostrar $\theta_{possivel}$ de uma distribuição auxiliar $q(\theta_{possivel} \mid \theta_{actual})$ "proposta". Esta distribuição pode ser simétrica, ou seja, $q(\theta_i \mid \theta_j) = q(\theta_j \mid \theta_i)$. O valor amostrado $\theta_{possivel}$ está sempre dependente do valor de θ_{actual}. Portanto o novo valor simulado está sempre dependente do valor imediatamente anterior, o que é uma característica das cadeias de Markov.
4. Calcular o rácio r = $\frac{\pi(\theta_{possivel}|y)}{\pi(\theta_{actual}|y)}$.
5. Amostrar μ com distribuição $U(0,1)$.
6. Aceitar como próximo ponto $\theta_{actual} = \theta_{possivel}$ se $\mu < min(r,1)$ caso contrário θ_{actual} fica inalterado.
7. Voltar a 3 e prosseguir iterativamente.

Sempre que $\pi(\theta_{possivel} \mid y) \geq \pi(\theta_{actual} \mid y)$ o novo valor $\theta_{possivel}$ é aceite, o que indica que os valores que *maximizam* a densidade $\pi(\theta \mid y)$ nunca são rejeitados; caso contrário, o novo valor também pode ser aceite, mas com probabilidade r. Verifica-se aqui a analogia com o Método de Rejeição, em que pontos *pertencentes* a zonas onde densidade da distribuição proposta é mais elevada são aceites em maior proporção. Os valores aceites terão um *comportamento* semelhante ao "gráfico" da densidade de $\pi(\theta \mid y)$ (Gelman et. al. 2003).

No que respeita ao rácio r, este pode ser escrito da seguinte forma,

$$r = \frac{\pi(\theta_{possivel} \mid y)}{\pi(\theta_{actual} \mid y)} = \frac{\frac{p(y|\theta_{possivel})\pi(\theta_{possivel})}{p(y)}}{\frac{p(y|\theta_{actual})\pi(\theta_{actual})}{p(y)}} = \frac{p(y \mid \theta_{possivel})\pi(\theta_{possivel})}{p(y \mid \theta_{actual})\pi(\theta_{actual})},$$

removendo a necessidade de integração mencionada em (3.4) para obter *a posteriori*. Assim, este método permite obter uma amostra da *posteriori* $\pi(\theta \mid y)$ sem ter a sua forma analítica, recorrendo apenas à verosimilhança e *a priori*.

O algoritmo Metropolis-Hastings é uma generalização do algoritmo Metropolis em que a distribuição auxiliar $q(\theta_{possivel} \mid \theta_{actual})$ não tem de ser simétrica. Aqui o rácio r tem de ser modificado para corrigir a assimetria, pelo que é calculado da seguinte forma:

$$r = \frac{\frac{p(y|\theta_{possivel})\pi(\theta_{possivel})}{q(\theta_{possivel}|\theta_{actual})}}{\frac{p(y|\theta_{actual})\pi(\theta_{actual})}{q(\theta_{actual}|\theta_{possivel})}} = \frac{p(y \mid \theta_{possivel})\pi(\theta_{possivel})q(\theta_{actual} \mid \theta_{possivel})}{p(y \mid \theta_{actual})\pi(\theta_{actual})q(\theta_{possivel} \mid \theta_{actual})}.$$

Uma vantagem deste método é que a possibilidade de haver transições assimétricas, o permitirá aumentar a velocidade do passeio aleatório (Gelman et. al. 2003).

3.1.5.3 Método de Amostragem de Gibbs

Este algoritmo pode ser considerado como um caso particular do algoritmo Metropolis-Hastings em que a distribuição auxiliar $q(\theta^{'} \mid \theta^{t})$ é a distribuição condicional do parâmetro θ_i dado os restantes elementos do vector de parâmetros θ, $\pi(\theta_i \mid \theta_{-i}, y)$ onde $\theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n)$.
A distribuição condicional de apenas um elemento do vector θ tomando como fixo os restantes elementos torna-se habitualmente mais fácil de amostrar dado que se trata de uma distribuição *univariada* (Gelman et. al. 2003).

O Algoritmo de Gibbs pode ser descrito da seguinte forma:

- Seleccionar $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$ o vector de parâmetros inicial arbitrário.
- De $j = 1$ até d actualizar o valor de θ_j através de $\theta_j \sim \pi(\theta_j \mid \theta_{-j}, y)$ de forma a que:
  - $\theta_1^{(1)} \sim \pi(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_d^{(0)}, y)$
  - $\theta_2^{(1)} \sim \pi(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_d^{(0)}, y)$
  - ...

– $\theta_j^{(1)} \sim \pi(\theta_3 \mid \theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{j-1}^{(1)}, \theta_{j+1}^{(0)}, ..., \theta_d^{(0)}, y)$

– ...

– $\theta_d^{(1)} \sim \pi(\theta_d \mid \theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{d-1}^{(1)}, y)$

Após finalizar este ciclo obtém-se $\theta^{(1)} = (\theta_1^{(1)}, ..., \theta_d^{(1)})$ a partir de $\theta^{(0)}$. Repetindo iterativamente este processo em que se gera $\theta^{(j)}$ através de $\theta^{(j-1)}$ obtém-se o seguinte conjunto de vectores:

- $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$
- $\theta^{(1)} = (\theta_1^{(1)}, ..., \theta_d^{(1)})$
- ...
- $\theta^{(t)} = (\theta_1^{(t)}, ..., \theta_d^{(t)})$

Resultados teóricos suportam que quando $t \to \infty$ o conjunto de vectores $(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_d^{(t)})$ tende em distribuição para o vector aleatório cuja função densidade conjunta é $\pi(\theta \mid y)$ (Paulino et. al. 2003), ou seja, os valores gerados convergem para a distribuição que pretendemos avaliar, obtendo-se assim uma amostra dos parâmetros para análise.

3.1.5.4 Convergência e Diagnóstico das Cadeias de Markov

Após obter uma amostra através da metodologia *MCMC* teremos necessariamente de verificar se as cadeias das quais os valores foram gerados atingiram o estado de equilíbrio. Para tal, existe um conjunto de métodos que permitem avaliar os valores obtidos de forma a *testar* se a cadeia, ou cadeias, atingiram uma situação de estacionariedade. Esta verificação é muito importante para poder fazer uma análise fiável. Serão seguidamente enunciados alguns dos métodos habitualmente usados e que se encontram em software de análise de cadeias *MCMC* (BRUGS, BOA, CODA). Maior detalhe técnico pode ser encontrado na documentação dos *packages* BOA, CODA e BRUGS cujos *websites* se encontram na bibliografia deste trabalho.

- Inspecção visual das cadeias

  Se for efectuada uma representação gráfica da série de dados gerados pela cadeia *MCMC* em condições de estacionariedade é esperado que o comportamento seja semelhante ao de um passeio aleatório, ou seja, uma sequência de "saltos" aleatórios que não apresentam um padrão perceptível.

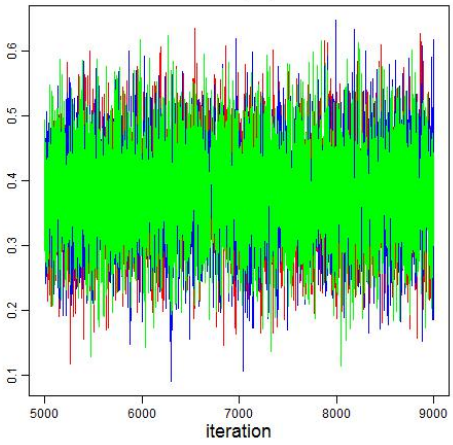


Figura 3.4: Exemplo do traço de três cadeias com comportamento aleatório

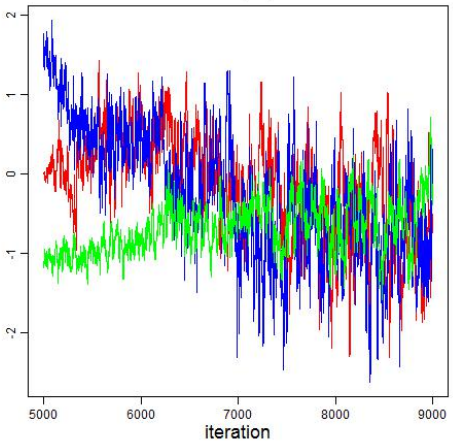


Figura 3.5: Exemplo do traço de três cadeias em que se verifica um comportamento não aleatório

- Geweke (1992)

  Trata-se de um método de diagnóstico efectuado para cada cadeia, sendo analisada a convergência da média ou de uma sua função. A cadeia é "partida" em duas partes que contêm aproximadamente 10% do início e 50% do final da cadeia. Se a cadeia for estacionaria as médias das duas partes da cadeia deverão ser similares. O diagnóstico é realizado através do teste à diferença das médias dos "pedaços" da cadeia. A diferença é dividida pelo seu desvio padrão, obtendo-se a estatística de teste Z (*Z score*). Quando $n \to \infty$ então Z tem assimptoticamente distribuição *N(0,1)*. Assim, valores da estatística Z pertencentes às caudas da distribuição normal padrão são indicativos de não estacionariedade.

- Gelman and Rubin (1992)

  Este é um teste diagnóstico para um conjunto de cadeias paralelas, iniciadas em pontos distintos. É usada informação sobre a variância das cadeias de modo a diagnosticar-se convergência. Suponha-se que foram geradas m cadeias de dimensão n. Para cada parâmetro de θ temos o conjunto de valores gerados $\theta_{ij(i=1,..,n;j=1,...,m)}$.

$$Seja \quad B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta_{.j}} - \bar{\theta_{..}})^2 \quad com \quad \bar{\theta_{.j}} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij} \quad e \quad \bar{\theta_{..}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta_{.j}}$$

$$e \quad seja \quad W = \frac{1}{m}\sum_{j=1}^{m}s_j^2 \quad com \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta_{.j}})^2.$$

  Estima-se a variância *a posteriori* $Var(\theta \mid y)$, através da média ponderada de B (variabilidade entre cadeias) e W (variabilidade dentro de cada cadeia) do seguinte modo:

$$\widehat{Var}(\theta \mid y) = \frac{n-1}{n}W + \frac{1}{n}B.$$

  Esta quantidade sobre-estima a variância da distribuição *posteriori*, sendo porém não enviesada quando $n \to \infty$. Por outro lado para um número finito de observações n, W subestima $Var(\theta \mid y)$, mas o seu valor esperado converge para a variância da *posteriori* quando $n \to \infty$. A convergência da cadeia *MCMC* é analisada através do *potential scale reduction* estimado por

$$\hat{R} = \sqrt{\frac{\widehat{Var}^{+}(\theta \mid y)}{W}},$$

  que tende para 1 à medida de $n \to \infty$. Assim se $\hat{R}$ não estiver suficientemente próximo de 1 não se verifica convergência para os valores de interesse θ e dever-se-á continuar a simulação (Gelman et. al., 2003).

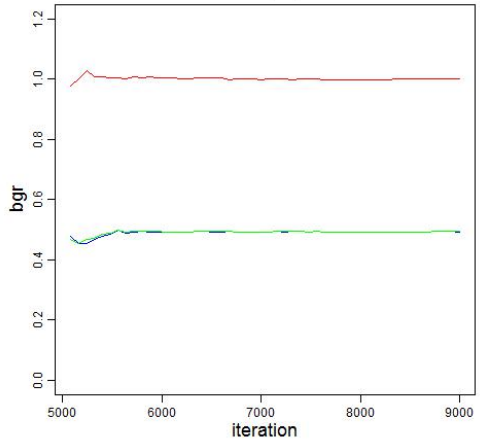


Figura 3.6: Gráfico Brooks-Gelman-Rubin (BRG) - exemplo de convergência de $\hat{R}$ para 1

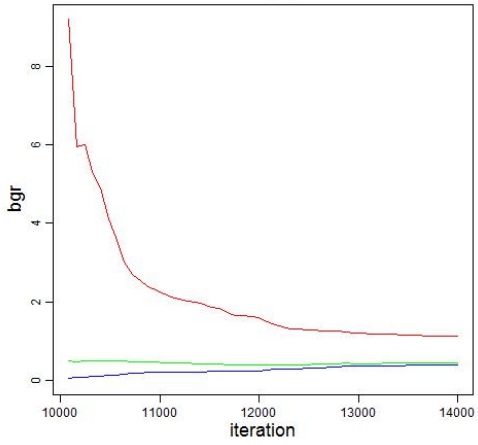


Figura 3.7: Gráfico Brooks-Gelman-Rubin (BRG) - exemplo de falha de convergência de $\hat{R}$ para 1

- Raftery and Lewis (1992)

  Um dos objectivos deste teste é obter um quantil específico q das distribuições simuladas, com um nível de erro r e probabilidade s. Após analisar a cadeia para os níveis q, r e s este procedimento calcula, para cada variável, o número de iterações necessárias até se "obter" convergência, o número de iterações que devem ser rejeitadas (*burn-in*) e o *thinning* adequado de forma obter uma amostra aproximadamente *i.i.d.*. É também calculado o *dependence factor*, que é um valor que nos indica o aumento relativo da amostra necessário para reduzir a autocorrelação.

- Heidelberg and Welch (1983)

  Este teste de convergência faz uso da estatística de teste Cramer-von-Mises para testar a hipótese nula H_o*: a amostra provém de uma distribuição estacionária.* O teste de Cramer-von-Mises é um teste não paramétrico usado para verificar se uma amostra $x_1, x_2, ..., x_n$ considerada i.i.d. tem uma função de distribuição $F(x)$ (contínua).

  No que respeita à analise da cadeia, o teste é aplicado ao total da cadeia, e caso a hipótese H_o seja rejeitada, o teste é feito sucessivamente a "pedaços" cada vez mais pequenos da cadeia. São removidos sucessivamente o equivalente a 10% do

total de cadeia até a hipótese H_o não ser rejeitada ou sobrar apenas 50% dos valores originais. Se se rejeitar H_o para a cadeia com os remanescentes 50% dos valores simulados é considerado que não foi atingida a estacionariedade e portanto terá de ser obtida uma cadeia mais *longa*.

- Autocorrelação

  Por definição, um valor de uma cadeia de *Markov* está dependente do elemento imediatamente anterior da respectiva cadeia. Neste sentido é natural existir autocorrelação na série de valores obtidos da cadeia de *Markov*. É portanto necessário verificar o nível de autocorrelação dos valores gerados, visto que se pretende obter uma amostra (aproximadamente) i.i.d. das distribuições que pretendemos analisar. São guardados diversos pontos da série de valores da cadeia, com determinado espaçamento entre eles, denominado *thinning*. Este distanciamento entre os valores amostrados permite a reduzir o nível de autocorrelação existente entre os valores gerados. Os restantes valores da cadeia, não seleccionados, são descartados da amostra.
  Para diagnosticar o nível de autocorrelação recorre-se à inspecção da função de autocorrelação (ACF) dos valores da cadeia gerada, o que permite escolher um *thinning* adequado de forma a reduzir a autocorrelação.

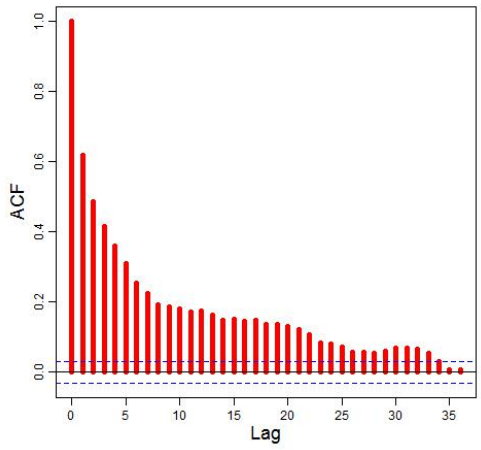


Figura 3.8: Exemplo de elevada autocorrelação - *thinning 0*

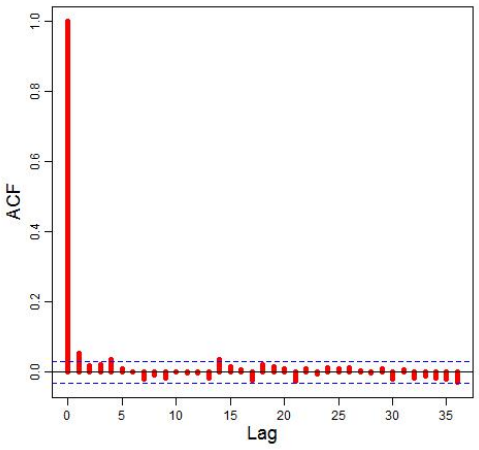


Figura 3.9: Exemplo de baixa autocorrelação - *thinning 35*

4

Definição de Modelos

4.1 Modelos

Nesta secção apresenta-se um resumo dos modelos que foram considerados na análise dos dados sobre a *schistosomose*. Serão detalhadas as suas principais características e a razão da sua aplicação a estes dados.

Modelo	**Zeros**	**Sobredispersão**	**Distribuição Truncada**
Poisson (P GLM)	Não	Não	Não
Poisson ZI (ZIP)	Sim	Não	Não
Poisson ZA (ZAP)	Sim	Não	Sim
Binomial Negativa (BN GLM)	Não	Sim	Não
Binomial Negativa ZI (ZIBN)	Sim	Sim	Não
Binomial Negativa ZA (ZABN)	Sim	Sim	Sim

Tabela 4.1: Resumo dos modelos utilizados

4.1.1 Poisson para dados de Contagens

Diz-se que uma variável aleatória Y tem distribuição de *Poisson* de parâmetro μ ($\mu > 0$) se a sua função de massa de probabilidade (f.m.p.) for dada por:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, ... \tag{4.1}$$

A variável aleatória Y tem valor médio e variância μ. Esta igualdade entre o valor

médio e a variância é uma característica marcante da distribuição de *Poisson* denominada *equidispersão*. Na prática esta particularidade não é frequentemente verificada, ocorrendo "problemas" de sobredispersão, sendo a variância explicada pelo modelo inferior à observada nos dados (Turkman, 2000 e Congdom, 2005). Esta situação pode ser ultrapassada usando modelos alternativos mais flexíveis ou reparametrizações do modelo Poisson. Estas alternativas são alvo de estudo neste trabalho, embora o modelo Poisson permaneça na análise como ponto de comparação ou referência.

4.1.1.1 Modelo de Regressão Poisson

Na modelação de variáveis de contagem no contexto dos *Modelos Lineares Generalizados* (GLM), é frequentemente usado o *Modelo Poisson Log-Linear*. Denomina-se log-linear devido à utilização da função logarítmica para *ligar* o valor médio da variável de contagem, que se assume ter uma distribuição de Poisson, e a componente linear do modelo de regressão. Desta forma consegue-se que a variável dependente ou resposta, que é uma função do modelo linear, seja sempre positiva, tal como se espera visto tratar-se de uma contagem. Esta função é utilizada em contexto dos *GLM* com a vantagem de que permite criar modelos multiplicativos e de interpretação simples.

Se considerarmos Y_i como sendo a variável em estudo que segue uma distribuição de *Poisson* e x_i= $(x_{i1}, x_{i2}, ..., x_{im})$ um conjunto de m covariáveis observadas relativamente ao indivíduo i, o *modelo Poisson Log-linear* pode ser resumido a :

$$Y_i \sim Poisson(\mu_i) \qquad com \qquad log(\mu_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \tag{4.2}$$

onde β_j, j=1,2,...,m são os coeficientes das covariáveis do modelo de regressão.

4.1.1.2 Modelo de Regressão Poisson *Zero Inflated* (ZIP)

Os modelos *Zero Inflated* (ZI) são usados para modelar variáveis de contagem que apresentam demasiados zeros. São compostos por uma mistura de um ponto de *massa* em *zero*, combinada com uma distribuição que permite modelar variáveis de contagem, neste caso a distribuição de Poisson.

É criada assim uma estrutura em que os *zeros* têm duas origens, do fenómeno associado à infecção e a outros factores dos quais possam resultar o *excesso de zeros*.

O modelo ZIP tendo em conta a f.m.p. em (4.1) pode ser formalizado como:

$$P(Y_i = y_i) = \begin{cases} \phi_i + (1-\phi_i) \times e^{-\mu_i} & \text{se } y_i = 0 \\ (1-\phi_i) \times \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{se } y_i > 0 \end{cases} \tag{4.3}$$

onde ϕ_i é a probabilidade de observar um *zero falso* ou não associado ao processo que determina o aparecimento de ovos do parasita. Em contexto de regressão o valor médio μ_i será modelado tal como indicado em (4.2) e a componente ϕ_i será obtida através da regressão logística, que é a opção usual para modelar uma variável binária, neste caso a *existência* ou *não* de um *zero falso*. Para modelar a probabilidade ϕ_i utiliza-se a função *logit*,

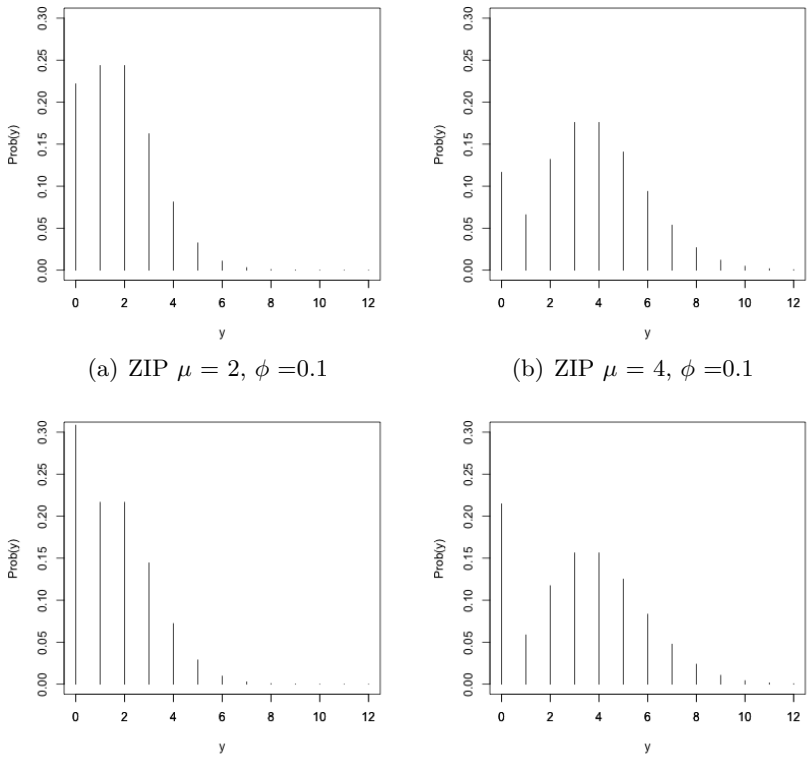
$$log\left(\frac{\phi_i}{1-\phi_i}\right) = \gamma_0 + \sum_{j=1}^{m} \gamma_j x_{ij} \tag{4.4}$$

que é equivalente a

$$\phi_i = \frac{e^{\gamma_0 + \sum_{j=1}^{m} \gamma_j x_{ij}}}{1 + e^{\gamma_0 + \sum_{j=1}^{m} \gamma_j x_{ij}}}.$$

As covariáveis usadas para modelar ϕ_i não têm de ser as mesmas que para μ_i, visto que estamos a modelar duas componentes distintas. De facto, os factores que influenciam o aparecimento de *zeros* em excesso não serão, à partida, os mesmos que influenciam a intensidade da infecção.

Este modelo tem valor médio $E(Y_i) = \mu_i\ (1-\phi_i)$ e variância $Var(Y_i) = (1-\phi_i)\ (\mu_i + \phi_i \cdot \mu_i^2)$. É simples verificar que a variância é sempre maior que o valor médio. Deste modo o modelo ZIP consegue transpor as circunstâncias restritivas do modelo Poisson no que diz respeito à dispersão e à sua adequabilidade às características dos dados.



(a) ZIP $\mu = 2$, ϕ =0.1 (b) ZIP $\mu = 4$, ϕ =0.1

(c) ZIP $\mu = 2$, ϕ =0.2 (d) ZIP $\mu = 4$, ϕ =0.2

Figura 4.1: Exemplos da f.m.p. do modelo ZIP

4.1.1.3 Modelo de Regressão Poisson *Zero Altered* (ZAP) ou de *Duas Partes*

Os modelos *Zero Altered* (ZA), de forma análoga aos ZI, são compostos por uma mistura em que há um ponto de *massa* em *zero* mas, neste caso, combinada com uma distribuição *truncada* em zero. Assim, o mecanismo que explica os *zeros* é único e o número de ovos do parasita observado é modelado por outra estrutura. Por esta razão considera-se que o modelo está dividido em duas partes.

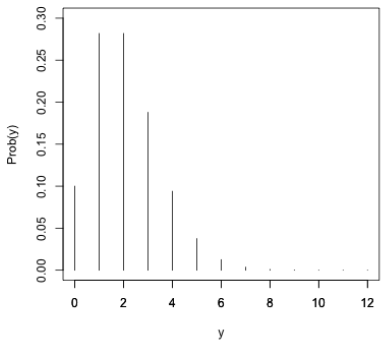
O modelo ZAP tem f.m.p. descrita da seguinte forma:

$$P(Y_i = y_i) = \begin{cases} \phi_i & \text{se } y_i = 0 \\ (1-\phi_i) \times \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!(1-e^{-\mu_i})} & \text{se } y_i > 0 \end{cases} \tag{4.5}$$

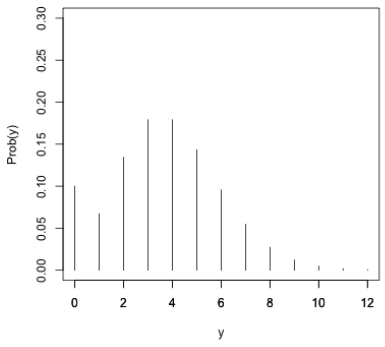
onde ϕ_i é a probabilidade de se observar um *zero* e $\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!(1-e^{-\mu_i})}$ é a f.m.p. da distribuição de Poisson truncada em zero. A origem dos zeros neste modelo é única e provém do ponto de massa ϕ_i. Para modelar μ_i e ϕ_i recorre-se à mesma metodologia de (4.2) e (4.3) respectivamente.

O modelo ZAP tem valor médio $E(Y_i) = \frac{1-\phi_i}{1-e^{-\mu_i}} \cdot \mu_i$ e variância $Var(Y_i) = \frac{1-\phi_i}{1-e^{-\mu_i}} \cdot (\mu_i + \mu_i^2) - \left(\frac{1-\phi_i}{1-e^{-\mu_i}} \cdot \mu_i\right)^2$.

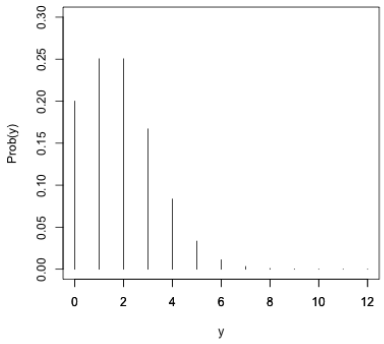
Se considerarmos $\psi_i = \frac{1-\pi_i}{1-e^{-\mu_i}}$ então o valor médio do modelo ZAP será $E(Y_i) = \psi_i \cdot \mu_i$ e a variância $Var(Y_i) = \psi_i \cdot \mu_i + \psi_i \mu_i^2 (1 - \psi_i)$. Fica visível que como ψ_i pode ser maior ou menor que 1, então este modelo pode acomodar-se tanto a sobredispersão como a subdispersão (Frees, 2009). Adicionalmente, o modelo ZAP pode ser considerado como uma reparametrização do modelo ZIP com $1 - \phi_i = (1 - \phi_i) \times (1 - e^{-\mu_i})$.



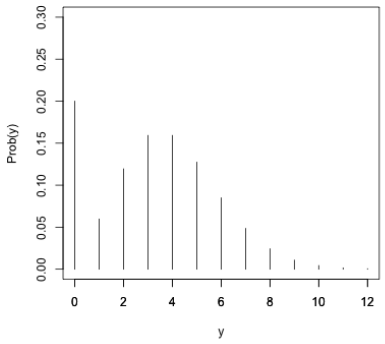
(a) ZAP $\mu = 2$, $\phi_{Altered}$ =0.1



(b) ZAP $\mu = 4$, $\phi_{Altered}$ =0.1



(c) ZAP $\mu = 2$, $\phi_{Altered}$ =0.2



(d) ZAP $\mu = 4$, $\phi_{Altered}$ =0.2

Figura 4.2: Exemplos da f.m.p. do modelo ZAP

4.1.2 Binomial Negativa para dados de Contagens

Uma variável aleatória Y tem uma distribuição *Binomial Negativa* se a sua *f.m.p.* for dada por:

$$P(Y = y) = \binom{k+y-1}{k-1} p^k (1-p)^y.$$

A variável Y é frequentemente designada como o número de tentativas falhadas até se conseguir k sucessos, sendo que p é a probabilidade de sucesso em cada tentativa. Esta distribuição também pode ser expressa na forma

$$P(Y = y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} p^k (1-p)^y . \tag{4.6}$$

Nesta expressão k pode ser considerado como um real, o que é conveniente para a modelação deste parâmetro, dito de *dispersão*, no modelo de regressão.

4.1.2.1 Modelo de Regressão Binomial Negativo

Apesar da simplicidade do modelo de regressão de Poisson, este é bastante restritivo no que respeita à igualdade da média e variância, característica que não se observa nos dados em análise neste trabalho. Em alternativa, para modelar contagens é usada a *Binomial Negativa*, com algumas vantagens sobre o modelo Poisson (Frees, 2011).
A *Binomial Negativa* tem dois parâmetros, o que permite maior flexibilidade que o modelo *Poisson*. Também, tendo em conta a f.m.p. em (4.6) e fazendo $p \to 1$ e $k \to 0$ de modo a que $k \cdot p \to \lambda$, prova-se que o modelo Poisson com valor médio λ é um caso limite do modelo *Binomial Negativo*, pelo que o modelo Poisson está encaixado no modelo *Binomial Negativo* (Cook, 2009). A distribuição *Binomial Negativa* também pode ser obtida de uma mistura de variáveis com distribuição *Poisson-Gama*, com a reparametrização do modelo *Poisson* de forma a que $\mu = \frac{k(1-p)}{p}$ (Cook, 2009).

Sendo Y $|\mu \sim$ Poisson$(\lambda\mu)$ com $\mu \sim Gama(\mu, k)$, prova-se que a distribuição marginal de Y é :

$$P(Y = y) = \int_0^{+\infty} p(y|\mu)p(\mu)d\mu \;=\; \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(1 - \frac{k}{k+\mu}\right)^y,$$

que não é mais que a f.m.p. da *Binomial Negativa* apresentada em (4.6), em que $p = \frac{k}{\mu+k}$, o valor médio é μ e a variância é $\mu(1 + \frac{\mu}{k})$. Com esta parametrização estamos em condições de modelar o valor μ de forma análoga a (4.2) onde k é mais um elemento

a estimar. Formalizando o modelo de regressão temos:

$$Y_i \sim Neg.Binomial\left(\frac{k}{k+\mu_i}, k\right) \qquad com \qquad log(\mu_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}, \tag{4.7}$$

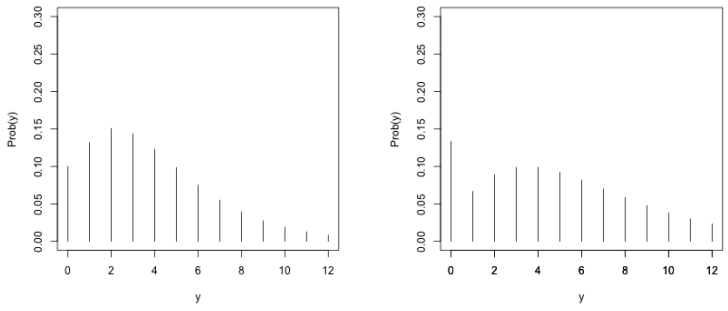
em que $\beta_{j,\ j=1,2,\ldots,m}$ são os coeficientes das covariáveis no modelo de regressão.

4.1.2.2 Modelo de Regressão Binomial Negativa *Zero Inflated* (ZIBN)

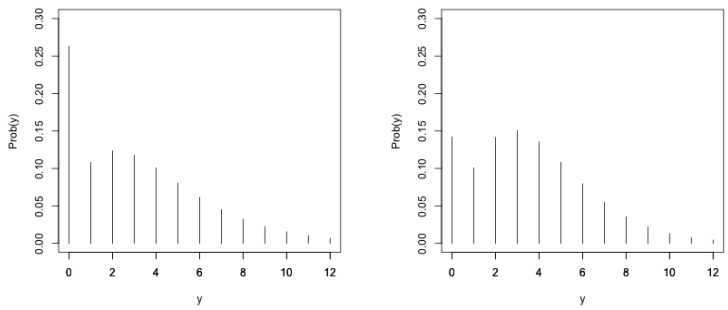
A construção do modelo *ZIBN* é análoga ao do *ZIP*. A sua f.m.p. pode ser descrita como:

$$P(Y_i = y_i) = \begin{cases} \phi_i + (1-\phi_i) \times \left(\frac{k}{k+\mu_i}\right)^k & \text{se } y_i = 0 \\ (1-\phi_i) \times \frac{\Gamma(y+k)}{y_i!\Gamma(k)} \left(\frac{k}{k+\mu_i}\right)^k \left(1 - \frac{k}{k+\mu_i}\right)^{y_i} & \text{se } y_i > 0 \end{cases} \tag{4.8}$$

onde ϕ_i é a probabilidade de observar um *zero falso* na observação i. O valor médio da variável de resposta μ_i será modelado como em (4.2) e a componente ϕ_i será obtida através de regressão logística como em (4.4). Este modelo tem valor médio $E(Y_i) = (1-\phi_i) \cdot \mu_i$ e variância $Var(Y_i) = (1-\phi_i)\left(\mu_i + \frac{\mu_i^2}{k}\right) + \mu_i^2 \cdot (\phi_i^2 + \phi_i)$.



(a) ZINB $\mu = 4$, ϕ =0.1, k=3 (b) ZINB $\mu = 6$, ϕ =0.1, k=3



(c) ZINB $\mu = 4$, ϕ =0.2, k=3 (d) ZINB $\mu = 4$, ϕ =0.1, k=6

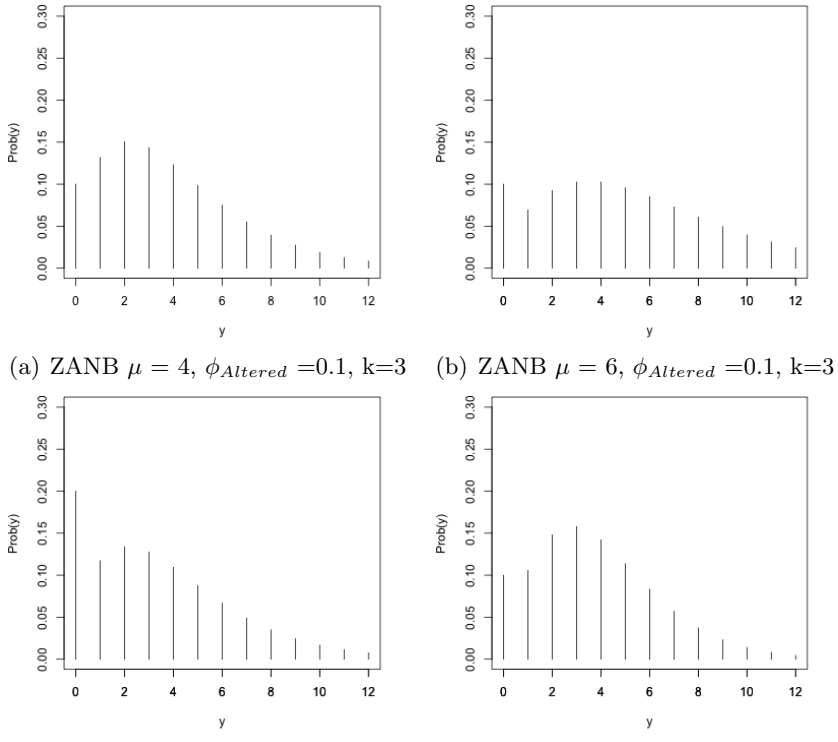
Figura 4.3: Exemplos da f.m.p. do modelo ZINB

4.1.2.3 Modelo de Regressão Binomial Negativa *Zero Altered* (ZABN) ou Modelo *Hurdle*

Tal como o modelo ZAP em (4.5) o modelo ZABN pode ser representado da seguinte forma:

$$P(Y_i = y_i) = \begin{cases} \phi_i & \text{se } y_i = 0 \\ (1-\phi_i) \times \frac{\Gamma(y_i+k)}{y_i!\Gamma(k)} \left(\frac{k}{k+\mu_i}\right)^k \left(1-\frac{k}{k+\mu_i}\right)^{y_i} / \left(1-\left(\frac{k}{k+\mu_i}\right)^k\right) & \text{se } y_i > 0 \end{cases} \tag{4.9}$$

onde ϕ_i é a probabilidade de observar um *zero* na observação i.
Este modelo tem valor médio $E(Y_i) = \frac{1-\phi_i}{1-P_0} \cdot \mu_i$ onde $P_0 = \left(\frac{k}{\mu_i+k}\right)^2$ e variância $Var(Y_i) = \frac{1-\phi_i}{1-P_0} \cdot \left(\mu_i^2 + \mu_i + \frac{\mu_i^2}{k}\right) - \left(\frac{1-\phi_i}{1-P_0} \cdot \mu_i\right)^2$. Também se trata de uma reparametrização do modelo *ZINB* de forma análoga ao que se verificou no modelo ZIP.



(a) ZANB $\mu = 4$, $\phi_{Altered}$ =0.1, k=3 (b) ZANB $\mu = 6$, $\phi_{Altered}$ =0.1, k=3

(c) ZANB $\mu = 4$, $\phi_{Altered}$ =0.2, k=3 (d) ZANB $\mu = 4$, $\phi_{Altered}$ =0.1, k=6

Figura 4.4: Exemplos da f.m.p. do modelo ZANB

4.1.3 Log-Verosimilhança dos Modelos

Depois de descrita a estrutura probabilística dos modelos podemos definir a verosimilhança de forma a prosseguir para a obtenção das estimativas *a posteriori*. Esta subsecção é um ponto de referência das log-verosimilanças dos modelos considerados. Será usada a habitual indexadação para cada indivíduo i, para um conjunto de dados de dimensão n e onde $I_\alpha(y)$ é a função indicatriz, definida por:

$$I_{(\alpha)}(y) = \begin{cases} 1 & se \quad y = \alpha \\ 0 & se \quad y \neq \alpha \end{cases}$$

Modelo
Poisson GLM (P GLM) $\sum_{i=1}^{n}[\ y_i\ log(\mu_i) - \mu_i - log(y_i!)\ ]$
Poisson ZI (ZIP) $\sum_{i=1}^{n} I_{(0)}(y_i)\ [log(p_i + (1-p_i)e^{-\mu_i})] +$ $\sum_{i=1}^{n}(1 - I_{(0)}(y_i))\ [log(1-p_i) - \mu_i + y_i log(\mu_i) - log(y_i!)]$
Poisson ZAP (ZAP) $\sum_{i=1}^{n}[\ I_{(0)}(y_i)\ log(p_i) + (1 - I_{(0)}(y_i))(log(1-p_i) + \mu_i + y_i log(\mu_i) - log(y_i!) - log(1-e^{-\mu_i})\ ]$
Binomial Negativa GLM (BN GLM) com reparametrização $\mu_i = \frac{k(1-p_i)}{p_i}$ e $r_i = \frac{k}{k+\mu_i}$ $\sum_{i=1}^{n}[\ log(\Gamma(y_i+k)) - log(y_i!) - log(\Gamma(k)) - k\ log\,(p_i) + y_i\ log\,(1-p_i)\ ]$
Binomial Negativa ZI (ZIBN) $\sum_{i=1}^{n} I_{(0)}(y_i)\ [\ log(p_i + (1-p_i)\ r_i^k)\ ] +$ $\sum_{i=1}^{n}(1 - I_{(0)}(y_i))[\ log(1-p_i) + log(\Gamma(y_i+k)) - log(y_i!) - log(\Gamma(k)) + k\ log\,(r_i) + y_i\ log\,(1-r_i)\ ]$
Binomial Negativa ZA (ZABN) $\sum_{i=1}^{n} I_{(0)}(y_i)\ log(p_i) +$ $\sum_{i=1}^{n}\ (1 - I_{(0)}(y_i))[\ log(1-p_i) + log(\Gamma(y_i+k)) - log(y_i!) - log(\Gamma(k)) + k\ log\,(r_i) + y_i\ log\,(1-r_i) -$ $log(1-r_i^k)\ ]$

Tabela 4.2: Descrição da função de Log-Verosimilhança dos modelos considerados

5

Dados e Covariáveis

Tal como referido anteriormente, os dados aqui tratados foram recolhidos pela Dr. Jacinta Teresa para a sua dissertação em Parasitologia Médica (Figueiredo, 2008) e já foram explorados por Olivença (2011). Desta forma, não se pretende neste capítulo aprofundar ou repetir análises descritivas realizadas noutras teses de mestrado. Pretende-se sim, de forma resumida, apresentar os dados, a metodologia de recolha e explorar o conjunto de covariáveis que foram consideradas neste trabalho. Para obter informação descritiva adicional, com grande detalhe, poder-se-á recorrer a Olivença (2011) e informação acerca da metodologia de recolha e análise *in situ* em Figueiredo (2008).

O dados foram obtidos mediante entrevista aos indivíduos com base na sua comparência ao inquérito. Foi explicado aos inquiridos o método de recolha das amostras e facultados recipientes próprios para a recolha da amostra de urina (Figueiredo, 2008).

Os indivíduos provêm de três províncias distintas, *Luanda, Bengo e Kwanza Sul*; o grupo em análise é constituído por elementos de ambos os sexos com idades entre os 15 e 75 anos. A maioria dos indivíduos provém de comunidades pobres com acesso restrito a água canalizada e instalações sanitárias. Dada a natureza da *schistosomose* a informação sobre o acesso e utilização da água é relevante. Devido à relação entre esta doença e problemas do sistema urogenital é também interessante, na perspectiva epidemiológica, avaliar a presença de sangue na urina ou hematúria, dado que esta pode indicar a existência de lesões provocadas pelo parasita (Figueiredo, 2008).
Também foi recolhida informação sobre o *motivo* e o *local de contacto habitual* com água, sendo este o principal *veículo* de dispersão ou propagação do parasita.

É também relevante a *naturalidade* dos indivíduos que participaram no inquérito já que muitos residentes na zona de recolha da amostra provêm de outras áreas nas quais a doença é endémica (Figueiredo, 2008), podendo assim terem transportado o parasita da sua localidade de origem.
Em Olivença (2011) verificou-se que existem frequências muito baixas de inquiridos em algumas províncias de origem e optou-se por agregar o conjunto de áreas originais em função da sua proximidade geográfica em 4 novos grupos. Esses mesmos agrupamentos foram utilizadas nesta análise.

Covariável	Zona Agregada	Zonas Iniciais
Naturalidade	Luanda, Bengo	Luanda e Bengo
	Bié, Huambo, Moxico	Bié, Huambo e Moxico
	Norte	Cabinda, Zaire, Uige, Kwanza Norte, Kwanza Sul e Malange
	Sul	Benguela e Huila

Tabela 5.1: Agregação da Zona de Naturalidade

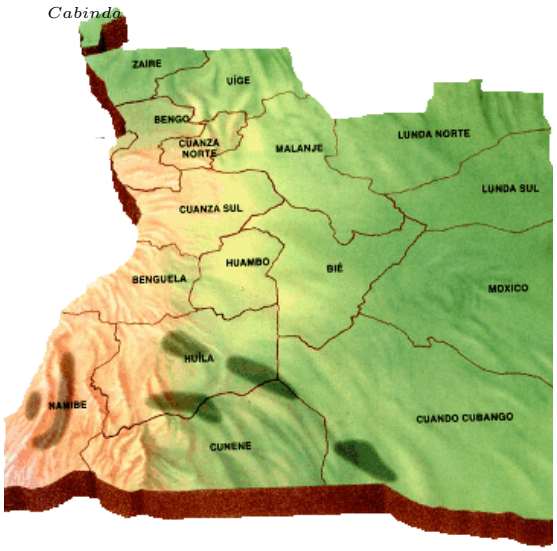


Figura 5.1: Mapa de Angola

.
Retirado e adaptado de *http://embangola.artedesign-net.pt/content.php?id=geografia*

5.1 Informação Descritiva

Nesta secção serão descritas algumas das informações mais relevantes acerca da informação recolhida e revisitada alguma parte do trabalho de Olivença (2011). Cada uma das covariáveis será analisada e representada graficamente. A distribuição do número de ovos por amostra de 10 ml de urina tem uma cauda direita *pesada*. De forma a dar melhor visibilidade a esta variável nas representações *gráficas*, a série de dados foi *censurada* de modo a só incluir indivíduos com uma contagem até um máximo de 100 ovos.

Em Olivença (2011) é indicado que dos 300 indivíduos que constituem a amostra, 85 não apresentam ovos no exame da amostra de 10 *ml* de urina, o que corresponde aproximadamente a 28,3% do total da amostra. Mais de um quarto das amostras de urina não apresentam ovos, o que indicia um *excesso de zeros* na distribuição da contagem de ovos por amostra.

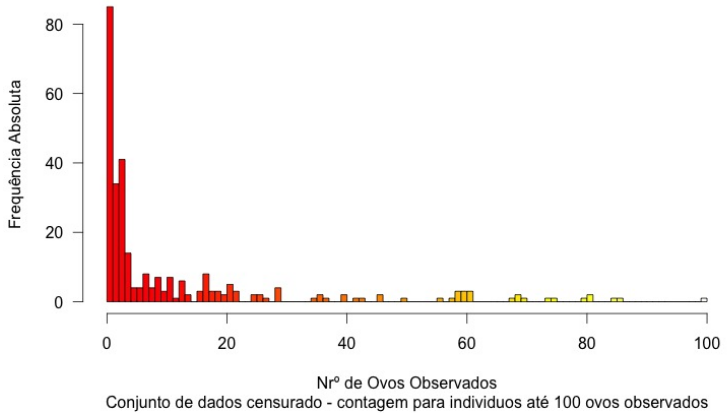


Figura 5.2: Diagrama de barras do número de ovos observado por amostra de urina

Amostra	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Com *zeros*	0	0	2	16	925	24,25	89.30
Sem *zeros*	1	2	7	21	925	33,84	104,00

Tabela 5.2: Estatísticas do número de ovos observado por amostra de urina

Nas estatísticas da Tabela 5.2 verifica-se que o desvio padrão é bem mais elevado que a média, mesmo para a amostra sem *zeros*, o que é indicativo de sobredispersão.

Também é visível no gráfico apresentado na Figura 5.2 uma distribuição de contagens com uma grande concentração de valores pequenos e uma cauda alongada, estendendo a distribuição para o lado direito, havendo valores elevados tendencialmente afastados da média.

No que respeita aos indivíduos com um número elevado de ovos, existem 33 indivíduos com 50 ou mais ovos observados na sua amostra de urina. É um conjunto caracterizado por indivíduos jovens, 19 (30%) indivíduos deste grupo tem menos de 25 anos. São maioritariamente estudantes (10 indivíduos, 30,3% do grupo), agricultores (8 indivíduos, 24,2% do grupo), trabalhadores domésticos (8 indivíduos, 24,2% do grupo). O principal motivo de contacto com a água deste subgrupo é recolher água em rios. Na sua generalidade, os elementos deste grupo não têm conhecimento da doença.

5.1.1 Covariáveis

Apresenta-se nesta subsecção uma análise descritiva do conjunto de covariáveis de forma a que se possa observar a sua relação com o número de ovos observado, a prevalência da infecção e avaliar com detalhe cada uma das covariáveis consideradas.

Na análise que se segue, os indivíduos nos quais se observaram ovos de *schistosoma* na respectiva amostra de urina foram considerados como "infectados". Desta forma a proporção de infectados será a fracção de indivíduos com ovos observados pelo número total de indivíduos. Note-se que um indivíduo pode estar infectado, mas não apresentar ovos do parasita na respectiva amostra de 10 ml de urina.

Idade

Verifica-se pela distribuição da *idade* que os indivíduos da amostra são maioritaria-

Covariável	Com ovos observados			Sem ovos observados	Total		
Intervalo de Idade	Nr^o de indivíduos	Nr^o Médio de ovos	Desvio Padrão	Nr^o de indivíduos	Nr^o de indivíduos	Proporção de Infectados	% de Indivíduos do Total
15-24	96	42	111,5	36	132	72,7 %	44,0 %
25-34	50	45,9	147,61	18	68	73,5 %	22,7 %
35-44	31	11,45	16,89	19	50	62,0 %	16,7 %
45-54	24	11,5	14,75	0	31	77,4 %	10,3 %
55-64	3	5,33	5,85	5	3	100,0 %	1,0 %
65-75	11	27,45	28,97	7	16	68,8 %	5,3 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.3: Estatísticas dos indivíduos por Idade

mente jovens, sendo que 66.7% dos inquiridos têm menos de 35 anos, o que é sintomático

de um país com esperança média de vida em 2012 de aproximadamente 52 anos (WHO 2012). Avaliando os valores da Tabela 5.3 observa-se haver uma tendência para indivíduos mais velhos apresentarem carga parasitária mais baixa. Porém, esta indicação deve ser encarada com cautela, dado o tamanho reduzido da amostra. Em Cardoso (2010) refere-se que a prevalência e a intensidade da infecção aumentam durante a infância e que posteriormente vão diminuindo com a idade. O mesmo trabalho sugere que indivíduos expostos à infecção acabam por desenvolver alguma resistência à reinfecção (isto baseado em resultados de inquéritos epidemiológicos).

Na representação gráfica do número de ovos em função da idade na Figura 5.3, a redução da intensidade da infecção à medida que a idade é mais avançada fica mais evidente.

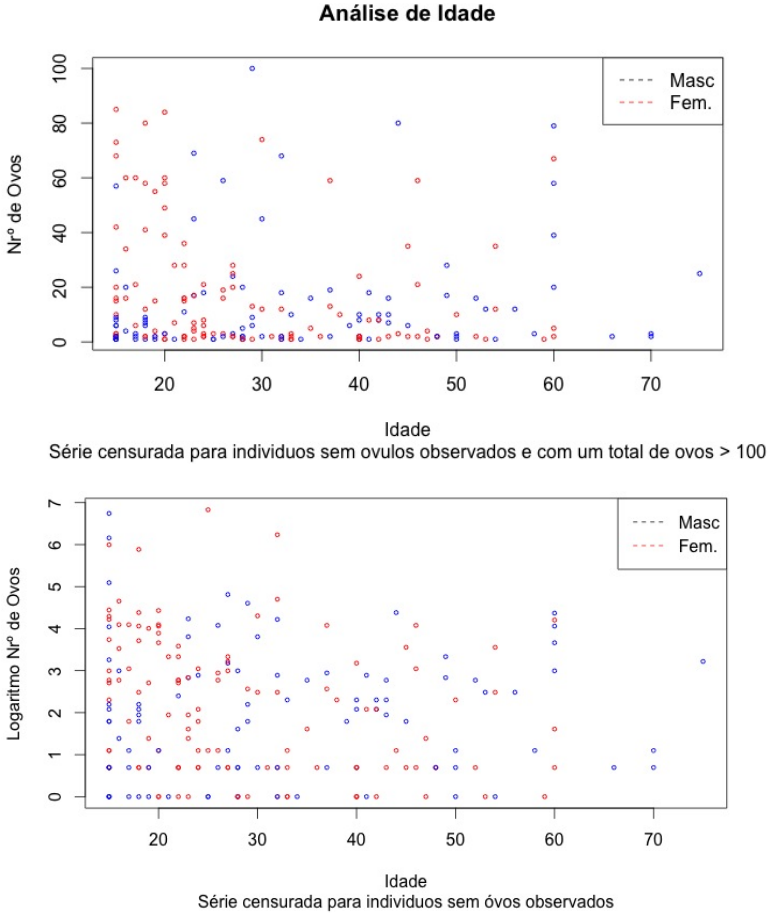


Figura 5.3: Número de ovos observado em função da Idade

Ambos os géneros aparentam uma diminuição da intensidade da infecção à medida que a *idade* avança, mas não fica perceptível, nesta representação gráfica, diferenças significativas em função do género dos indivíduos.

Género

Covariável	Com ovos observados			Sem ovos observados	Total		
Género	Nr° Indivíduos	Nr° Médio de ovos	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Masculino	94	30,68	101,17	40	134	70,1 %	44,6 %
Feminino	121	36,29	106,5	45	166	72,8 %	55,3 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.4: Estatísticas segundo o Género dos indivíduos

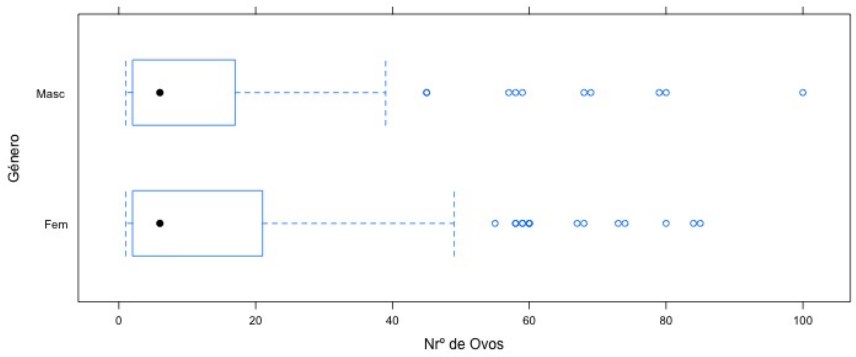


Figura 5.4: *Box-Plots* do número de ovos observado segundo o Género

Na amostra recolhida as mulheres têm maior representatividade, constituindo aproximadamente 55% do total de inquiridos. Pela análise da Tabela 5.4, os indivíduos do género feminino aparentam ter carga parasitária e prevalência de infecção ligeiramente mais elevada que os homens, sendo a diferença pequena. Nesta visão univariada, a covariável género parece ter pouca influência na explicação da variabilidade observada na *carga parasitária*. Testes estatísticos efectuados em Olivença (2011) sugerem que não existem diferenças significativas no número de ovos do parasita entre os indivíduos de diferentes géneros. Na bibliografia relativa a esta doença é indicado que existem diferenças significativas na intensidade entre homens e mulheres, o que se deve mais a diferenças económico-sociais entre os géneros do que a características genéticas,

biológicas ou imunológicas (Bruun et al. 2008). Em epidemiologia o género é considerado como indispensável na modelação, pelo que esta covariável será mantida nos modelos utilizados neste trabalho.

Província

Covariável	Com ovos observados			Sem ovos observados	Total		
Província de Residência	Nr° Indivíduos	Nr° Médio de ovos	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Luanda	99	35,78	111,19	47	146	67,8 %	48,6 %
Kuanza Sul	54	18,83	22,98	14	68	79,4 %	22,6 %
Bengo	62	43,8	131,6	24	86	72,0 %	28,6 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.5: Estatísticas segundo a Província de Residência

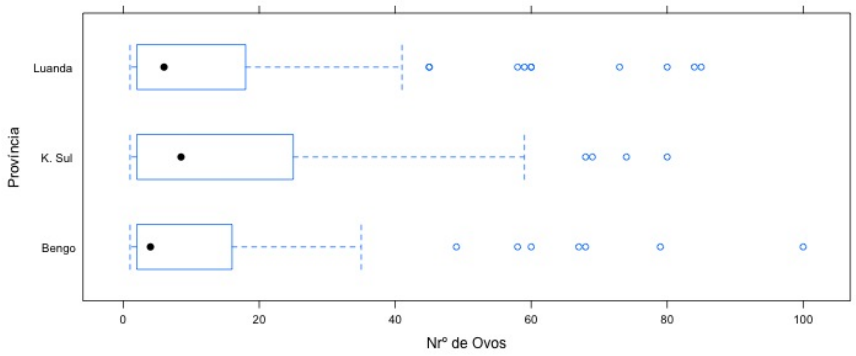


Figura 5.5: *Box-Plots* do número de ovos observado segundo a Residência

Quase metade dos inquiridos provém da província de Luanda. Nos dados recolhidos, Luanda apresenta a maior percentagem de não infectados. Em contrapartida, os residentes em Kuanza Sul, apesar de exibirem um número médio de ovos por amostra mais baixo, têm a maior prevalência de infecção das províncias amostradas.

Naturalidade

Covariável	Com ovos observados			Sem ovos observados	Total		
Naturalidade	Nr^o Indivíduos	Nr^o Médio de ovos	Desv. Padrão	Nr^o Indivíduos	Nr^o Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Sul	10	102,5	263,06	3	13	76,9 %	4,3 %
Norte	52	17,3	20,39	17	69	75,3 %	23,0 %
Luanda, Bengo	99	37,44	88,36	34	133	74,4 %	44,3 %
Bié, Huambo, Moxico	54	30,44	126,07	31	85	63,5 %	28,3 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.6: Estatísticas segundo a Naturalidade

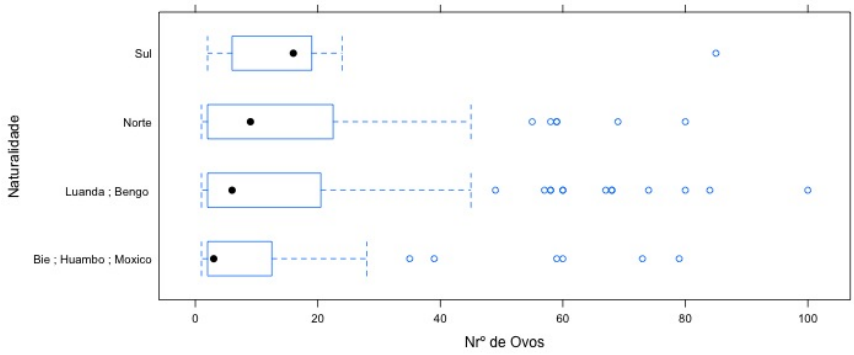


Figura 5.6: *Box-Plots* do número de ovos observado segundo a Naturalidade

Através da Tabela 5.6 é visível que cerca de metade da amostra é natural de Luanda - Bengo, o que é concordante com a informação acerca do local residência dos inquiridos. Os indivíduos da região Norte apresentam o mais baixo valor médio de ovos de parasita e os naturais de Bié, Huambo e Moxico a menor prevalência da doença.

Profissão

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Profissão	Nr° Indivíduos	Nr° Médio de ovos	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Trab. Doméstico	60	32,48	119,95	24	84	71,4 %	28,0 %
Outros	24	26,95	36,41	15	39	61,5 %	13,0 %
Funcio. Público	15	22,53	37,18	4	19	78,9 %	6,3 %
Estudante	35	77,91	176,11	11	46	76.1 %	15,3 %
Agricultor	81	19,93	58,6	31	112	72,3 %	37,3 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.7: Estatísticas segundo a Profissão

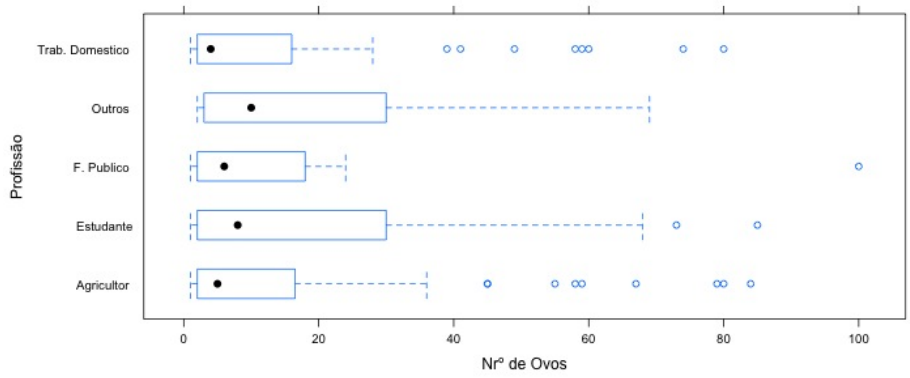


Figura 5.7: *Box-Plots* do número de ovos observado segundo a Profissão

A maior parte do indivíduos da amostra são agricultores, representando quase 40% da amostra, seguidos pelos *trabalhadores domésticos* com 28% do total de observados e pelos *estudantes* que representam 15% da amostra. Cerca de metade dos *agricultores* são homens. Já o grupo de *trabalhadores domésticos* é essencialmente constituído por mulheres.

Os inquiridos que indicam ser *agricultores* e que apresentam ovos na amostra de urina têm o número médio de ovos mais baixo de todas as profissões. Em Olivença (2011) é comentado que esta situação não é de esperar dado o contacto usual que estes indivíduos teriam com a águas não tratadas. Este grupo ocupacional é a segunda profissão com idade média mais elevada, o que poderá ser indicativo de um conjunto de indivíduos com maior resistência à doença. No extremo oposto, os *estudantes* são o grupo com valores mais elevados, algo que se justificará em parte pela juventude dos elementos

Profissão	Idade Média (anos)
Trab. Doméstico	28,9
Outros	37,8
Funcio. Público	29,4
Estudante	15,4
Agricultor	35,8
Total	30,6

Tabela 5.8: Idade média dos indivíduos por Profissão

que o constitui e que foram identificados como sendo particularmente susceptíveis à doença. A profissão que agrega o conjunto *outros* tem a menor taxa de infecção, mas é um conjunto com pouca expressão no grupo em análise, representando apenas 13% da amostra recolhida.

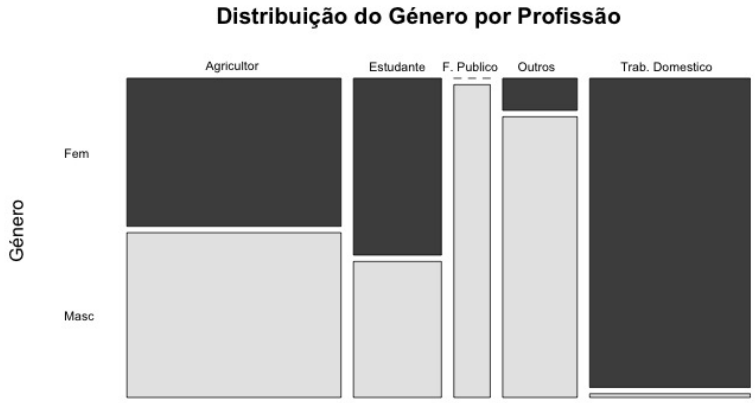


Figura 5.8: Distribuição dos indivíduos por Género e Profissão

Género	Agricultor	Estudante	F. Público	Outros	Trab. Doméstico	Total
Feminino	53	26	0	4	83	166
Masculino	59	20	19	35	1	134
Total	112	46	19	39	84	300

Tabela 5.9: Distribuição dos indivíduos por Género e Profissão

Na análise das profissões por género apresentado na Figura 5.8, é de referir que a maioria das mulheres no estudo indicam ser *agricultoras* ou *domésticas*, com a nota adicional de que cerca de metade dos *agricultores* são mulheres. No entanto, não existe

um único elemento cuja profissão seja *funcionário público* e seja mulher.

Conhecimento da Doença

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Conhecimento da doença	Nr° Indivíduos	Nr° Médio	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Não sabe	178	37,14	113,48	66	244	73,0 %	81,3 %
Sabe	37	17,97	25,93	19	56	66,1 %	18,7 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.10: Estatísticas segundo o Conhecimento da Doença

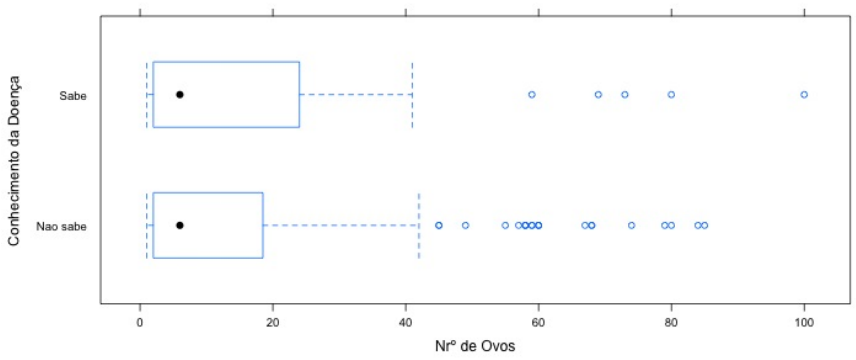


Figura 5.9: *Box-Plots* do número de ovos observado segundo o Conhecimento da Doença

Apenas um pequeno grupo de indivíduos indica ter conhecimento desta doença (18.7%). Estes apresentam um número médio de ovos significativamente mais baixo que os indivíduos que indicam não ter conhecimento, o que pode ser indício que este grupo de pessoas toma medidas preventivas contra a infecção e/ou já recebeu tratamento contra a doença anteriormente.

Conhecimento da Doença	Agricultor	Estudante	F. Público	Outros	Trab. Doméstico	Total
Não sabe	96	33	14	29	72	244
Sabe	16	13	5	10	12	56
Total	112	46	19	39	84	300

Tabela 5.11: Distribuição dos indivíduos em função do Conhecimento da Doença por Profissão

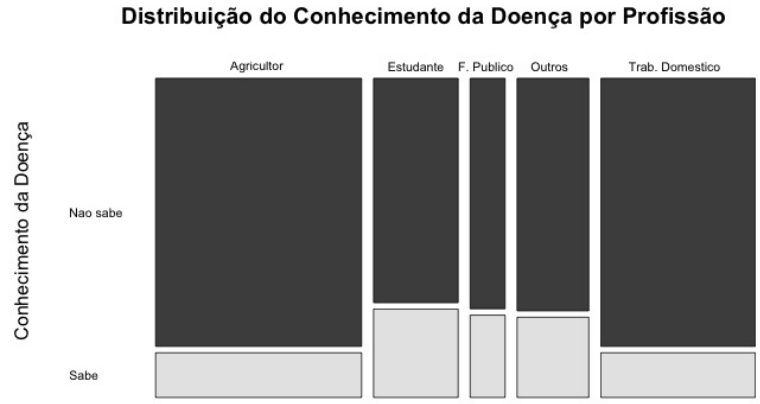


Figura 5.10: Distribuição do nível de Conhecimento da Doença por Profissão

Na Figura 5.10 é possível ver que o nível de desconhecimento acerca da doença é elevado entre os inquiridos, mesmo face ao elevado nível de prevalência desta doença, que nesta amostra é de 71.7%. Os *agricultores* e *trabalhadores domésticos* apresentam a maior proporção de pessoas que desconhece a *schistosomose*. No grupo dos *estudantes*, onde deve existir maior nível de informação, um pouco menos de um terço indica conhecer a doença. É relevante questionar se há diferenças na prevalência de infecção entre ter ou não conhecimento desta doença dentro do mesmo grupo ocupacional.

Profissão	Não tem conhecimento	Tem conhecimento
Trab. Doméstico	34,67	18,25
Outros	28,73	20,2
Funcio. Público	18,6	30,4
Estudante	96,37	15,62
Agricultor	21,05	12,81
Total	37,14	17,97

Tabela 5.12: Análise do número médio de ovos dos indivíduos infectados por Profissão em função do Conhecimento da Doença

Na Tabela 5.12, relativa unicamente a indivíduos *infectados*, verifica-se que dentro do mesmo grupo ocupacional, os indivíduos com conhecimento da doença apresentam cargas parasitárias mais baixas, excepto no caso dos *funcionários públicos*.

Teste Hematúria

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Resultado do Teste de Hematúria	Nr^o Indivíduos	Nr^o Médio de ovos	Desv. Padrão	Nr^o Indivíduos	Nr^o Indivíduos	Proporção de Infectados	% Indivíduos do Total
Negativo	161	21,31	75,89	77	238	67,7 %	79,3 %
Positivo	54	71,2	156,16	8	62	87,1 %	20,7 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.13: Estatísticas segundo os resultados do Teste Hematúria

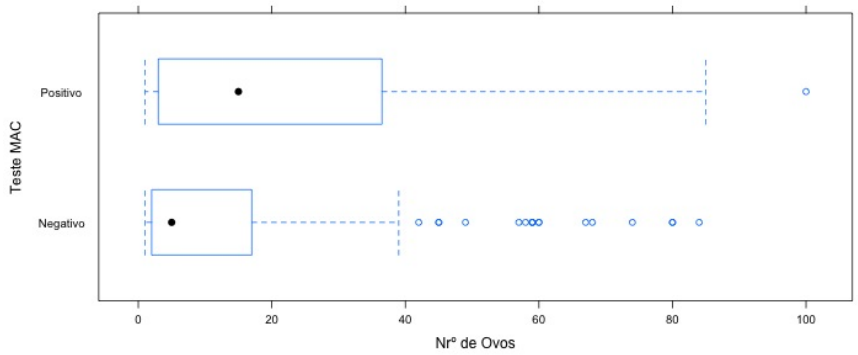


Figura 5.11: *Box-Plots* do número de ovos observado segundo o resultado do Teste Hematúria (ou de Hematúria)

Como seria de esperar, dada a associação entre a *schistosomose* e lesões urogenitais, quando o teste de hematúria é positivo o número médio de ovos é mais elevado e a respectiva proporção de *infectados* é também mais alta (87,1% de infectados no caso de teste positivo contra 67,7% no caso de resultado negativo). Portanto, o resultado do teste de hematúria será um bom indicador da existência e intensidade da doença.

Verifica-se que dos 62 elementos com teste positivo, 40 têm entre 14 e 24 anos, o que é de esperar, visto que os mais jovens são um dos principais grupos de risco.

Local de Contacto com a Água

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Local de Contacto	Nr° Indivíduos	Nr° Médio de ovos	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Tanque	52	15,42	27,16	31	83	62,6 %	27,6 %
Rio	142	42,02	125,8	37	179	79,3 %	59,6 %
Lagoa	21	24,14	28,82	17	38	55,2 %	12,6 %
Total	215	33,84	104	85	300	71,6 %	100 %

Tabela 5.14: Estatísticas segundo o Local de Contacto com a Água

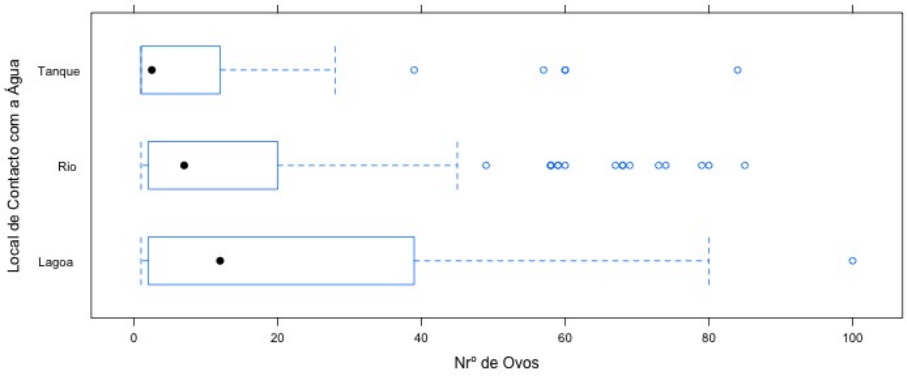


Figura 5.12: *Box-Plots* do número de ovos observado segundo o Local de Contacto com a Água

Quase 60% dos inquiridos dizem ter contacto com *água do rio* e apenas um quinto destes não apresenta ovos na contagem da amostra de urina. Dos que dizem ter contacto com *água de lagoas*, quase 45% aparentam ainda não estar infectados, sendo o *local* com a mais baixa prevalência da infecção. Os inquiridos que indicam ter contacto com *água dos rios* apresentam uma prevalência e intensidade média da infecção mais elevada; já o grupo que indica ter contacto com *água de tanque* apresenta um número médio de ovos francamente mais baixo que os restantes indivíduos. Supõe-se que a água dos tanques seja proveniente das chuvas, embora investigadores afirmem que em muitos casos a água dos tanques é recolhida dos rios e daí o aparecimento de casos de infecção (Olivença, 2011).

Motivo de Contacto com a Água

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Motivo de Contacto	Nr° Indivíduos	Nr° Médio de ovos	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% Indivíduos do Total
Todos os Outros	16	14,43	22,21	4	20	80,0 %	6,6 %
Pescar	30	32,9	86,09	9	39	76,9 %	13,0 %
Nadar	10	111,7	263,83	7	17	58,8 %	5,7 %
Lav. Roupa	22	19,9	27,78	2	24	91,7 %	8,0 %
Higiene Pessoal	28	12,85	16,93	10	38	73,7 %	12,7 %
Buscar Agua	109	38	112,09	53	162	67,3 %	54,0 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.15: Estatísticas segundo o Motivo de Contacto com a Água

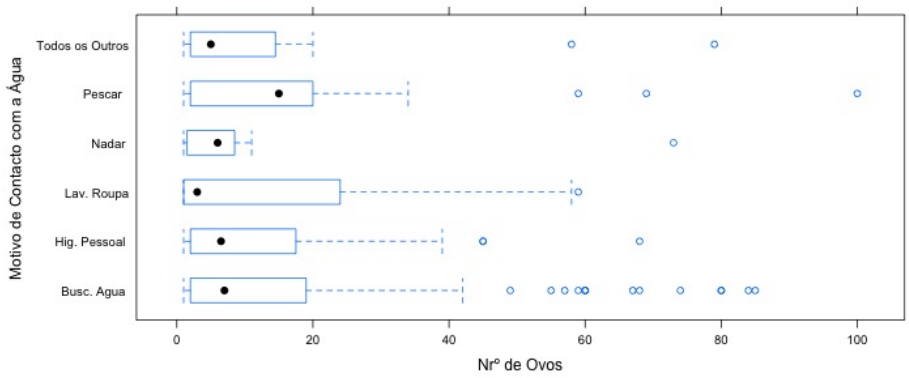


Figura 5.13: *Box-Plots* do número de ovos observado segundo o Motivo de Contacto com a Água

O motivo dominante para o contacto é *buscar água*, com mais de metade dos indivíduos da amostra a fornecerem essa indicação. Esta situação é natural em populações com difícil acesso a água canalizada e que naturalmente terão de estar concentradas em locais em proximidade de colecções de água, como rios ou lagos. De facto, dos 162 indivíduos cujo principal motivo de contacto é *buscar água*, 119 são *agricultores* ou *trabalhadores domésticos*, o que indica que, para uma elevada proporção da amostra, são as suas actividades laborais que as colocam em risco de infecção. O contacto com a água, tanto devido à actividade laboral como por outras necessidades, é determinante na exposição ao parasita mas também para a sua dispersão.

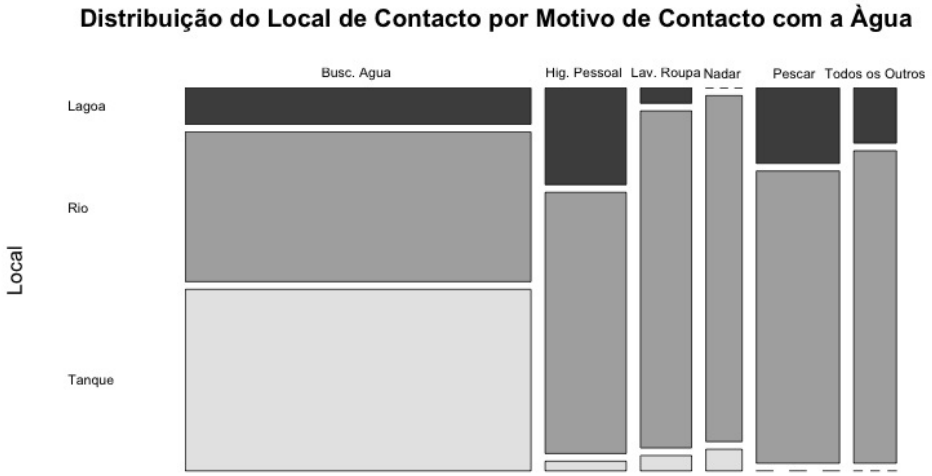


Figura 5.14: Distribuição dos indivíduos por Local de Contacto segundo o Motivo de Contacto com Água

Local	Buscar Agua	Higiéne Pessoal	Lavar Roupa	Nadar	Pescar	Todos os Outros	Total
Lagoa	16	10	1	0	8	3	38
Rio	66	27	22	16	31	17	179
Tanque	80	1	1	1	0	0	83
Total	162	38	24	17	39	20	300

Tabela 5.16: Distribuição dos indivíduos por Local de Contacto segundo o Motivo de Contacto com Água

As diferentes actividades decorrem sobretudo nos *rios*, que são um dos principais *pontos* de disseminação do parasita. Também se verifica que um dos locais mais comuns para para ir buscar água são os *tanques*. Os indivíduos que indicam ter contacto com a água em *tanques* apresentam a mais baixa taxa de infecção dos diferentes locais considerados, mas tal como se verificou anteriormente, também poderão ser um foco de disseminação do parasita.

Existência de Água Canalizada

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Existência de Água Canalizada	Nr° Indivíduos	Nr° Médio de ovos	Desv. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Tem	14	28,5	46,27	7	21	66,6 %	7 %
Não tem	201	34,21	106,92	78	279	72,0 %	93 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.17: Estatísticas segundo a Existência de Água Canalizada

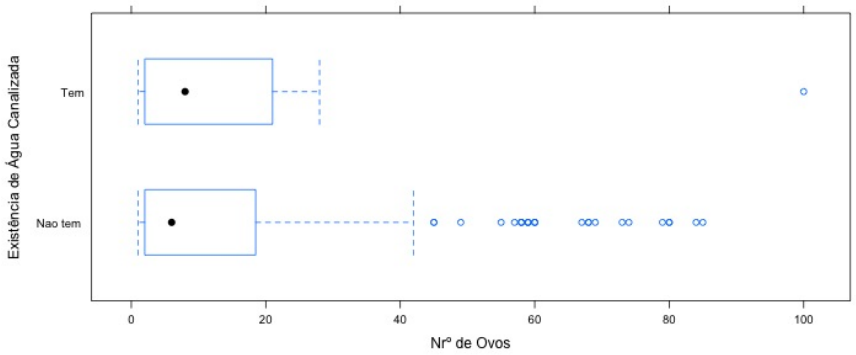


Figura 5.15: *Box-Plots* do número de ovos observado segundo a Existência de Água Canalizada

A maioria dos indivíduos da amostra indica que não tem acesso a *água canalizada*, apenas 7% dos indivíduos da amostra indica ter água canalizada. Há indícios de que ter *água canalizada* será benéfico, a avaliar pelo número médio de ovos do parasita mais reduzido quando é reportada a existência de água canalizada e também pela menor prevalência da infeccção que se verifica na Tabela 5.17. Um dos principais motivos indicados para a elevada prevalência desta doença é o acesso restrito a "fontes" de água segura ou tratadas (King, 2010). A reduzida proporção dos inquiridos que refere ter água canalizada, o que é indicativo da necessidade dos indivíduos de se deslocarem a rios ou lagos para aceder a reservas de água, potencialmente expondo-se à infecção de forma repetida.

Existência de WC dentro de casa

Covariável	Com Ovos Observados			Sem Ovos Observados	Total		
Existência de WC	Nr° Indivíduos	Nr° Médio	Des. Padrão	Nr° Indivíduos	Nr° Indivíduos	Proporção de Infectados	% de Indivíduos do Total
Fora de Casa	166	39,33	117,16	66	232	71,5 %	77,3 %
Dentro	49	15,24	24,11	19	68	72,0 %	22,7 %
Total	215	33,84	104	85	300	71,7 %	100 %

Tabela 5.18: Estatísticas segundo a Existência de WC Dentro de Casa

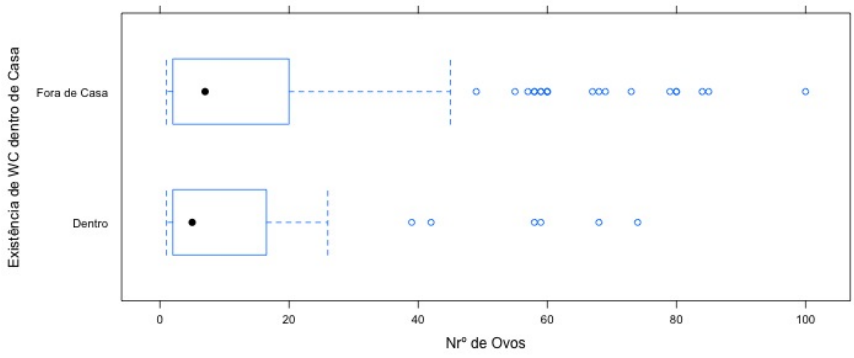


Figura 5.16: *Box-Plots* do número de ovos observado segundo a Existência de WC dentro de Casa

Apenas um quinto dos indivíduos da amostra afirma ter *wc* dentro de casa, apresentando um número médio de ovos mais baixo do que os que referem não ter. A proporção de infectados que se observa da Tabela 5.18 não indica que haja grande diferença na prevalência da doença entre ter ou não *wc* dentro de casa. Isto à partida poderá parecer contraditório, mas no contexto em que se inserem os inquiridos, o *wc* é um simples vaso sanitário ou latrina, não oferecendo grandes melhorias no que respeita a condições sanitárias.

A ausência de saneamento básico tem como consequência a contaminação de colecções da água doce através dos excrementos humanos que contêm ovos de *schistosoma*. Comparando os valores médios de ovos na Tabela 5.15, a existência de *wc* em casa aparenta oferecer alguma protecção contra o parasita através do controlo das excreções humanas, visto que a intensidade da infecção aparenta ser mais baixa.

6

Resultados dos Modelos

Foi utilizado o package BRUGS do R para obter amostras da distribuição *posteriori* dos parâmetros dos modelos através da metodologia *MCMC*. Este package invoca o programa OpenBugs. Para cada modelo foram usadas várias cadeias e analisada a convergência das mesmas.
O package BRUGS utiliza também o package CODA de forma a permitir facilmente, e através da linha de comando do R, analisar a convergência das cadeias geradas. Assim, todo o trabalho está organizado em torno do R e o respectivo código será disponibilizado em anexo conjuntamente com esta dissertação.

6.1 Definição das distribuições *a Priori*

Nos diversos modelos foram utilizadas distribuições *a priori* vagas ou pouco informativas. Para cada parâmetro dos preditores lineares dos modelos foi utilizada como *priori* a distribuição Normal, centrada em zero e com variância muito elevada ($\mathcal{N}(0, 1/0.0001)$). No caso dos modelos que utilizam a distribuição Binomial Negativa, para o parâmetro de dispersão, e tendo em conta que este é positivo, foi utilizada como *priori* a distribuição Gamma, com parâmetro de forma pequeno e de escala elevado de forma a introduzir um mínimo de informação e adequar a *priori* aos valores admissíveis (Gamma(0.0001, 1/0.0001)). As variâncias elevadas das distribuições utilizadas permitem que a informação dos dados domine o resultado final, o que reflecte o facto de que partimos na ausência de informação *a priori*.

6.2 Resultados

Nesta secção pretende-se mostrar o resultado final dos diferentes modelos, as variáveis seleccionadas e fazer comparação entre os modelos.
Os modelos foram ajustados com todas as covariáveis e sucessivamente removidas as covariáveis consideradas não significativas. Após a análise da amostra da distribuição *a posteriori* de cada parâmetro dos modelos utilizou-se a seguinte regra: se o valor zero estivesse dentro do Intervalo de Máxima Densidade de Probabilidade (HPD) a 95%, também denominado como Intervalo de Máxima Densidade (HDI) (Krush, 2010), esta covariável seria candidata a ser removida do modelo. Seguidamente seria analisado o modelo sem esta covariável. Em caso de melhoria em termos dos indicadores Akaike's Information Criterion (AIC) e Deviance Information Criterion (DIC) face ao modelo anterior, a covariável seria removida, caso contrário seria mantida no modelo. Este processo prossegue iterativamente, removendo covariáveis até só remanescerem as consideradas significativas.
No caso das covariáveis com vários níveis, em que um ou vários níveis poderiam ser considerados não significativos, foi ajustado o modelo com e sem essa covariável e utilizado o critério anterior para decidir manter ou remover a covariável do modelo.

O indicador DIC, que consta da Tabela 6.1, foi proposto por Spiegelhalter et al. (2002) e é definido como a soma do valor esperado *a posteriori* da deviance $\mathrm{D}(\theta) = -2Log(L(\theta \mid y))$ acrescido uma penalização devido à complexidade do modelo, algo semelhante aos indicadores AIC ou Bayesian Information Criterion (BIC). Neste caso a penalização é a diferença entre o valor esperado da deviance ($\overline{D(\theta)}$) e o valor da deviance calculado com o valor esperado das *posterioris* de θ ($D(\bar{\theta})$). Assim este indicador pode ser obtido da seguinte forma:

$$DIC(\theta) = \overline{D(\theta)} + (\overline{D(\theta)} - D(\bar{\theta})) = 2\overline{D(\theta)} - D(\bar{\theta}).$$

Em Spiegelhalter et al. (2002) é indicado que para modelos com pouca informação *a priori* (ou informação negligenciável) este indicador será aproximado do AIC.

Covariáveis	GLM	GLM	ZIP		ZIBN		ZAP		ZANB	
	Poisson	Binomial Neg.	Zeros	Média	Zeros	Média	Zeros	Média	Zeros	Média
Género	√	√	√	√	√	√	√	√	√	√
Idade	√	√	√	√	√	√	√	√	√	√
Água Canalizada										
Existência de WC	√			√		√		√		√
Contacto com a Água	√	√	√	√		√	√	√	√	
Conhecimento da Doença	√	√	√	√		√	√	√		√
Teste Hematúria	√	√	√	√		√	√	√	√	√
Profissão	√			√				√		
Motivo de Contacto	√	√		√		√		√		
Província	√		√	√			√	√	√	
Naturalidade	√			√				√		
$D(\theta)$	17596,1	1992,2	13897,5		2018,2		13893,9		2028,4	
$\overline{D(\theta)}$	17620	2005	13930		2002		13920		2042	
DIC	17643,8	2017,7	13962,4		1985,7		13946		2055,5	
AIC	17640,1	2018,2	13957,5		2052,2		13953,9		2058,4	
BIC	17847,1	2140,5	14239,7		2212,1		14236,1		2199,5	
Nr^o Parâmetros	22	13	30		17		30		15	
burn-in	5000	10000	20000		50000		10000		10000	
Tamanho da Amostra	4000	5000	4000		20000		4000		4000	
thinning	35	175	180		500		200		125	
Nr^o de Cadeias	3	3	3		2		2		3	

√ Covariável seleccionada

Tabela 6.1: Resumo dos Modelos

Na Tabela 6.1 verifica-se que os modelos que utilizam a distribuição Binomial Negativa apresentam melhores resultados nos vários indicadores calculados (AIC, DIC e BIC) e são mais parcimoniosos. A covariável que indica a existência ou não de água canalizada não foi considerada importante por nenhum modelo, o que se antevia pela análise descritiva. O conjunto de indivíduos que indicam ter água canalizada é reduzido e não se observaram diferenças relevantes entre ter ou não água canalizada. Em todos os modelos foram mantidas as variáveis idade e género dos indivíduos. No entanto, a idade mostrou ser um factor pouco importante como se poderá verificar na Tabela 6.2.

Dos modelos considerados, o modelo BN GLM é o mais parcimonioso de todos com 13 parâmetros e com os menores valores do AIC e DIC, quando comparados com as restantes alternativas. Em comparação, o modelo ZIBN apresenta resultados muito semelhantes com mais 4 parâmetros, mas com a vantagem de ter uma estrutura especializada para explicar o excesso de *zeros*.

No conjunto de modelos com recurso à distribuição Poisson, verifica-se uma melhoria em todos os indicadores nos modelos ZIP e ZAP em relação ao modelo Poisson GLM, o que indica que as estruturas específicas para modelar o excesso de *zeros* são eficazes e ajustam-se às características dos dados.

Adicionalmente, e no que respeita à afirmação sobre a semelhança entre os valores de

DIC e AIC referida em em Spiegelhalter et al. (2002), nestes modelos, nos quais foram utilizadas *prioris* não informativas, verificam-se diferenças reduzidas entre o DIC e o AIC.

Seguidamente apresenta-se a tabela com as médias das distribuições *a posteriori* dos parâmetros dos modelos.

Covariáveis	GLM	GLM	ZIP		ZIBN		ZAP		ZANB	
	Poisson	Bin. Neg.	Zeros	Média	Zeros	Média	Zero	Média	Zeros	Média
Ordenada na Origem	1,927	3,183	-0,915	1,772	-238,1	2,652	-0,977	1,765	-0,983	2,676
Género(F)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Género (M)	0,723	0,398	0,155	0,878	-29,54	0,406	0,207	0,882	0,203	-0,196
Idade	-0,019	-0,017	0,001	-0,014	-6,471	-0,017	0,004	-0,014	0,004	-0,021
Água Canal. (Sim)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Água Canal. (Não)	—	—	—	—	—	—	—	—	—	—
WC (Dentro)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
WC (Fora)	0,823	—	—	0,946	—	0,527	—	0,949	—	0,711
Cont. com água (Rio)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Cont. com água (Lagoa)	-0,365	-0,559	1,837	0,163	—	-0,467	1,881	0,164	1,872	—
Cont. com água (Tanque)	-1,525	-1,292	0,374	-1,628	—	-1,263	0,369	-1,63	0,375	—
Conhec. da doença (Sim)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Conhec. da doença (Não)	1,142	0,758	—	1,269	—	0,887	—	1,273	—	0,629
Hematúria (Negativo)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Hematúria (Positivo)	1,323	1,349	-1,233	1,116	—	1,296	-1,295	1,118	-1,283	0,998
Prof. (Func. Público)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Prof. (Estudante)	0,146	—	—	0,362	—	—	—	0,363	—	—
Prof. (Agricultor)	-0,428	—	—	-0,235	—	—	—	-0,235	—	—
Prof. (Trab. doméstico)	0,382	—	—	0,689	—	—	—	0,691	—	—
Prof. (Outras)	-0,095	—	—	0,256	—	—	—	0,257	—	—
Motivo (Buscar água)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Motivo(Higiene Pessoal)	-0,697	-0,89	—	-0,902	—	-0,888	—	-0,904	—	—
Motivo (Lavar roupa)	-1,45	-0,747	—	-1,755	—	-0,653	—	-1,761	—	—
Motivo (Pescar)	-1,974	-1,421	—	-2,246	—	-1,399	—	-2,255	—	—
Motivo (Nadar)	-0,44	-0,622	—	-0,258	—	-0,682	—	-0,258	—	—
Motivo (Todos os anteriores)	-1,633	-1,667	—	-1,62	—	-1,615	—	-1,963	—	—
Província (Luanda)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Província (Bengo)	0,089	—	-0,314	-0,112	—	—	-0,412	-0,114	-0,402	—
Província (Kwanza Sul)	0,351	—	-1,46	0,166	—	—	-1,407	0,167	-1,403	—
Naturalidade (Luanda Bengo)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Naturalidade (Bié Huambo Moxico)	0,354	—	—	0,609	—	—	—	0,611	—	—
Naturalidade (Norte)	-0,496	—	—	-0,607	—	—	—	-0,609	—	—
Naturalidade (Sul)	0,54	—	—	0,313	—	—	—	0,311	—	—
Parâmetro de dispersão: r	—	0,282	—	—	—	0,285	—	—	—	0,509

[a)] Categoria de Referência
— Nível da covariável não utilizado

Tabela 6.2: Médias a *posteriori* dos parâmetros dos modelos

Contrariamente ao que se tinha observado na parte descritiva das covariáveis, a covariável género parece ter influência nos resultados finais. Tal está de acordo com o que é conhecido acerca da *schistosomose*, de que há diferenças na infecção entre géneros. No entanto, esta situação não se verifica em todos os modelos, apenas nos modelos Poisson GLM, ZIP e ZAP (excepto nas estruturas que modelam os *zeros*) se

exclui a possibilidade deste parâmetro ser nulo.

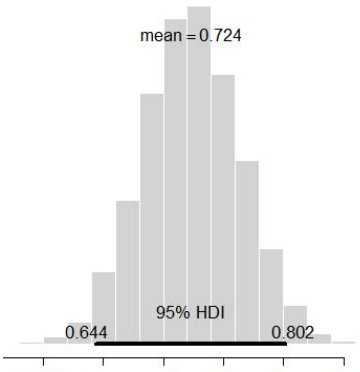


Figura 6.1: *Posteriori* da Variável Género: Masculino - Poisson GLM

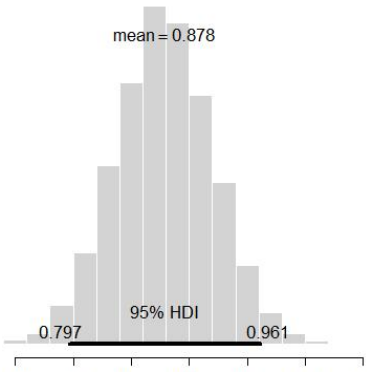


Figura 6.2: *Posteriori* da Variável Género: Masculino - ZIP

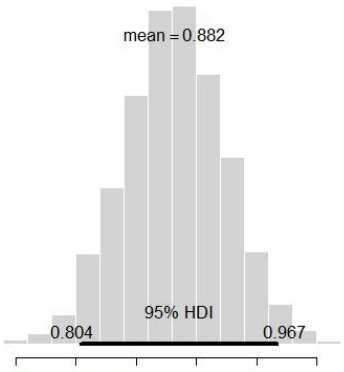


Figura 6.3: *Posteriori* da Variável Género: Masculino - ZAP

A covariável *idade*, mantida em todos os modelos, tem um impacto negligenciável na generalidade dos modelos. É possível verificar na distribuições *a posteriori* desta covariável, para os diferentes modelos, médias praticamente nulas e intervalos de máxima densidade de reduzida amplitude que contêm zero. De referir que nos diferentes modelos, no que respeita à modelação da resposta média, a idade apresenta "sinal negativo", o que está de acordo com a análise prévia em que à medida que a idade aumenta há uma ligeira redução da intensidade da infecção.

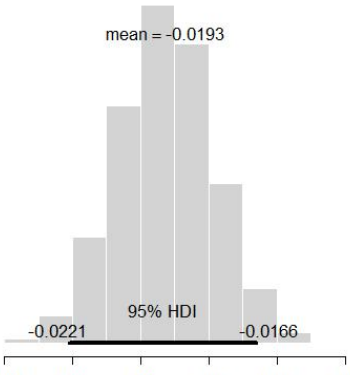


Figura 6.4: *Posteriori* da Variável Idade - Poisson GLM

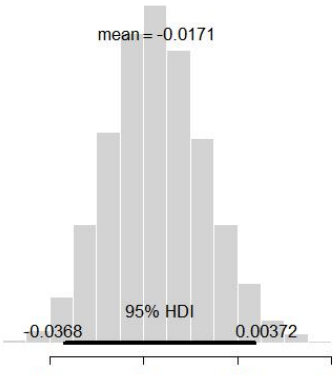


Figura 6.5: *Posteriori* da Variável Idade - Binomial Negativa GLM

Para o modelo Poisson GLM para o parâmetro associado à covariável *idade* o intervalo HDI é -0.00221 até -0.00166 e para o modelo Binomial Negativa GLM o intervalo é -0.0368 até 0.00372. Excepto o modelo Binomial Negativa GLM, todos os restantes

modelos usam a covariável que indica existência de *wc* dentro de casa. Esta é utilizada para modelar o número médio de ovos, não sendo considerada importante nos modelos com estruturas específicas para explicar o *excesso de zeros*.

A covariável que indica o resultado do teste de Hematúria também é usada em todos os modelos. É uma covariável com "peso" dados os valores elevados das distribuições *a posteriori* associadas a esta covariável. Repare-se que nos modelos ZI e ZA, quando esta covariável é usada para modelar os *zeros* tem valor negativo, o que indicará que um teste Hematúria positivo diminui a probabilidade de se observar um *zero* e quando é usada para prever o valor médio de ovos o valor é positivo, aumentado o número esperado de ovos.

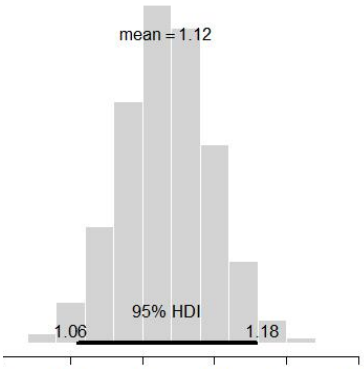


Figura 6.6: *Posteriori* da Variável Hematúria: Teste Positivo - ZIBN - Médias

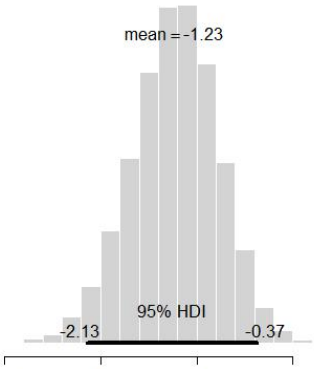


Figura 6.7: *Posteriori* da Variável Hematúria: Teste Positivo - ZIBN - Zeros

Também usada por todos os modelos, excepto nas estruturas do modelo logístico, é a informação sobre o conhecimento prévio da doença, no sentido em que o conhecimento da doença terá um efeito positivo, ou seja, os indivíduos com conhecimento da doença poderão recorrem a medidad preventivas, o que diminui a intensidade da infecção ou já terão sido alvo de tratamento previamente. No *website* da WHO, é possível encontrar um panfleto informativo com o nome *What children should know about Bilharzia* [1] com indicações sobre comportamentos a adoptar de forma diminui o risco de infecção.

[1] Bilharzia é outra denominação usada para a *schistosomose*

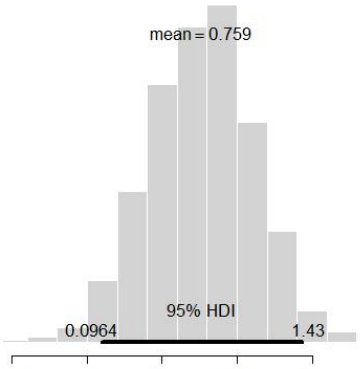


Figura 6.8: *Posteriori* da Variável Tem Conhecimento da Doença: Não Tem - Binomial Negativa GLM

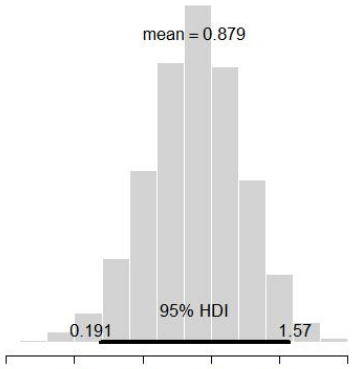


Figura 6.9: *Posteriori* da Variável Tem Conhecimento da Doença: Não Tem - ZIBN

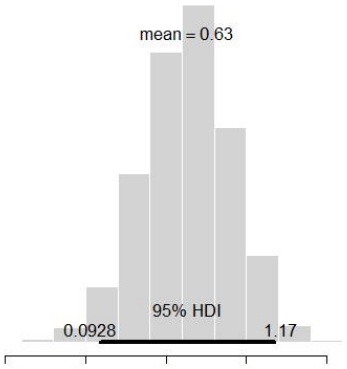


Figura 6.10: *Posteriori* da Variável Tem Conhecimento da Doença: Não Tem - ZABN

Outras covariáveis consideradas importantes na generalidade dos modelos são o *motivo de contacto* e a *localização* da água, o que é natural dado que a infecção dá-se justamente em contacto com água. Avaliando os valores da Tabela 6.2 para estas covariáveis, ir buscar água, principalmente em rios, tem um efeito prejudicial. Já a utilização de água de tanques, aparenta ser uma forma de mitigar a infecção.

No modelo ZIBN, todos os parâmetros do modelo logístico (Ordenada na origem, Género e Idade) associado à "inflação" de zeros têm valores bastantes elevados, sendo de questionar o grau de confiança em relação a estes parâmetros.

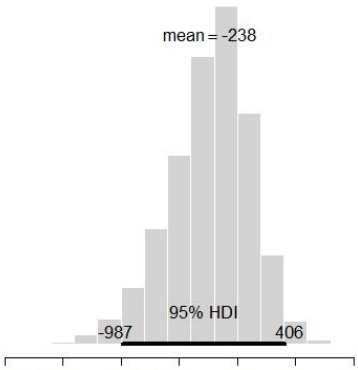


Figura 6.11: *Posteriori* da Ordenada na Origem - ZIBN - Zeros

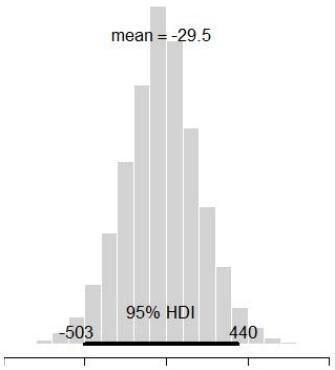


Figura 6.12: *Posteriori* da Variável Género: Masculino - ZIBN - Zeros

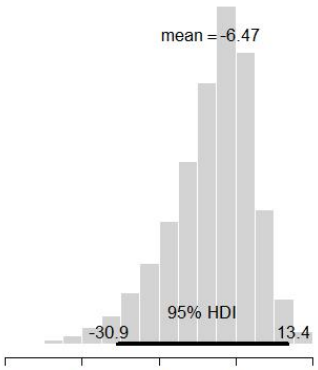


Figura 6.13: *Posteriori* da Variável Idade - ZIBN - Zeros

Nas Figuras 6.11, 6.12 e 6.13 fica clara a grande amplitude dos intervalos de máxima densidade. Como se pode observar não se exclui a possibilidade destes parâmetros serem nulos, o que revela grande incerteza em relação a estes parâmetros.

Na Tabela 6.1 é possível verificar a elevada dimensão da amostra usada para o modelo ZIBN, isto deve-se ao elevado erro *MC* que se verificou nas amostras obtidas. Aumentou-se 5 vezes o tamanho da amostra inicial, obtendo-se uma melhoria marginal em termos de erro *MCMC*. Devido aos elevados níveis de autocorrelação que se observam nas cadeias geradas para este modelo, é muito dispendioso em termos de tempo obter as distribuições *a posteriori*, não sendo praticável prosseguir com a simulação de amostras de maior dimensão.

O modelo ZIBN, sem a estrutura que modela o *excesso de zeros* reduz-se ao modelo Binomial Negativo GLM, o qual como se verifica na Tabela 6.1 tem resultados semelhantes ao ZIBN em termos de AIC ou DIC, sendo porém um modelo mais parcimonioso. Na análise *clássica* em Olivença (2011) estes valores também são negativos e elevados, o que é indicativo que o valor da função *logit* é muito pequeno, ou seja, a probabilidade de ser um *zero* ser falso é bastante reduzida. Neste sentido, a estrutura que modela o *excesso de zeros* tem pouco "peso" no modelo ZIBN. O que explica a grande semelhança entre o modelo Binomial Negativo GLM e o ZIBN.

6.2.1 Ordenadas Preditivas Condicionais

No caso em análise, com um conjunto de dados reduzido, não é conveniente particionar a amostra de forma ter um grupo de treino e outro para validação dos modelos. É proposto nestas situações utilizar a técnica *"leave one out"* recorrendo ao cálculo das *ordenadas preditivas condicionais* (CPO) dos modelos como forma de teste à sua capacidade preditiva.

Considerando y_1, y_2, y_3,... a amostra de dados em análise e θ o vector de parâmetros do modelo, a CPO para a observação i será:

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta, y_{-i})\pi(\theta \mid y_{-i})d\theta,$$

onde y_i é o elemento i da amostra e y_{-i} representa o conjunto de elementos da amostra excepto y_i. A CPO de y_i representa a probabilidade *a posteriori* de se observar y_i, considerando que o modelo foi ajustado com todos os valores, excepto a observação y_i. As CPO são uma forma de avaliar o quão provável é uma determinada observação dada a informação das restantes. Valores baixos para as ordenadas preditivas condicionais

são indicação de falta de ajustamento (Paulino et. al., 2003).
Os valores das CPO podem ser estimadas através das cadeias *MCMC* geradas, a estimativa é obtida do seguinte modo (Ntzoufras, 2009):

$$\widehat{CPO} = \frac{1}{T^{-1}\sum_{t=1}^{T} p(y_i \mid \theta^{(t)})}.$$

Trata-se da média harmónica da função de massa (ou densidade) de probabilidade de y_i para o conjunto de elementos da amostra de dimensão T, gerados após o respectivo período de *burn-in*. A demonstração deste "resultado" pode ser encontrada em Paulino et. al. (2003) pág. 354.
Verificou-se que para vários indivíduos, nos modelos com distribuição de Poisson, o denominador das estimativas das CPO seria zero ou aproximadamente zero (≈ 0), o que não permite obter as estimativas das respectivas CPO. De forma a ultrapassar este entrave somou-se um valor pequeno (0.0001) a cada $p(y_i \mid \theta^{(t)})$ de forma a ser possível obter estes dados. Nos gráficos que se seguem os valores mais "extremos" das CPO correspondem a esses indivíduos. Esta situação deve-se às características restritivas do modelo Poisson no que respeita à dispersão, deste modo para os indivíduos em que o valor médio estimado é distante do valor observado as estimativas das CPO são muito baixas ou nulas. Apresenta-se agora gráficos dos valores das CPO contra o número de ovos de parasita observado para cada indivíduo, com o objectivo de avaliar o comportamento dos modelos em função da intensidade da infecção.

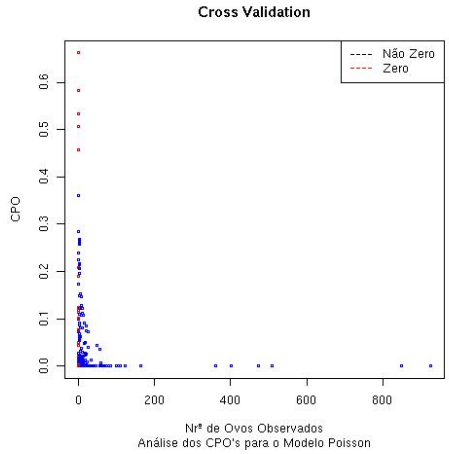


Figura 6.14: CPO Poisson por número de ovos observado por indivíduo

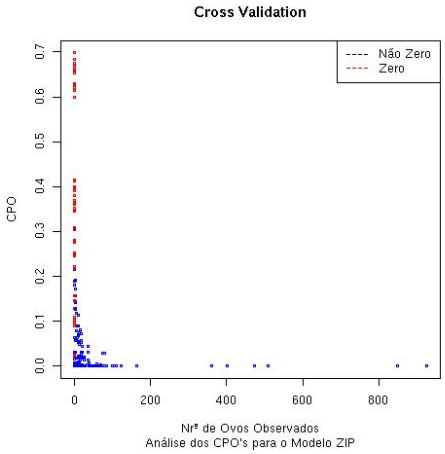


Figura 6.15: CPO ZIP por número de ovos observado por indivíduo

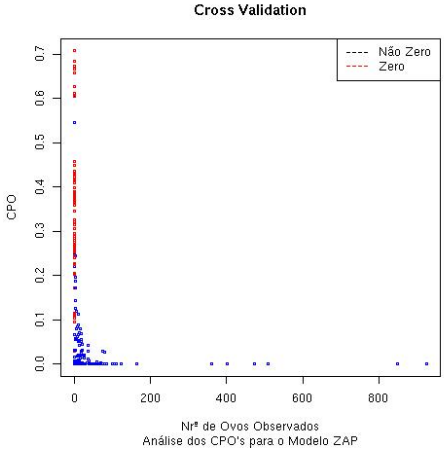


Figura 6.16: CPO ZAP por número de ovos observado por indivíduo

Os modelos com a distribuição de Poisson apresentam valores bastante baixos das CPO para indivíduos com um número elevado de ovos de parasita. Nos gráficos das Figuras 6.14, 6.15 e 6.16 nota-se um decréscimo muito rápido das CPO à medida que os valores de ovos observados aumentam, o que é indicativo de limitações dos modelos em prever valores elevados de intensidade de infecção. As CPO do modelo ZAP representadas na Figura 6.16, apresentam um grande número de pontos vermelhos com valores elevados, o que indica uma elevada capacidade de prever *zeros* face ao outros modelos.

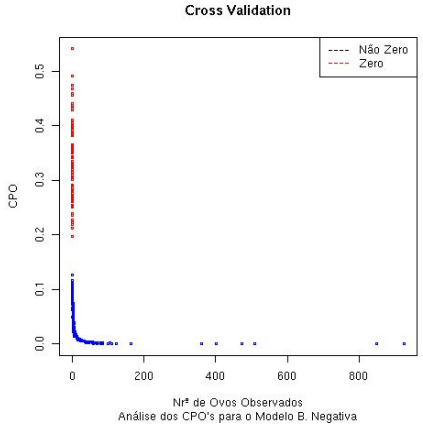


Figura 6.17: CPO Binomial Negativa GLM por número de ovos observado por indivíduo

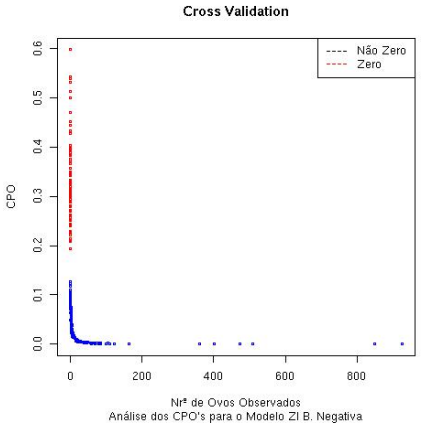


Figura 6.18: CPO ZIBN por número de ovos observado por indivíduo

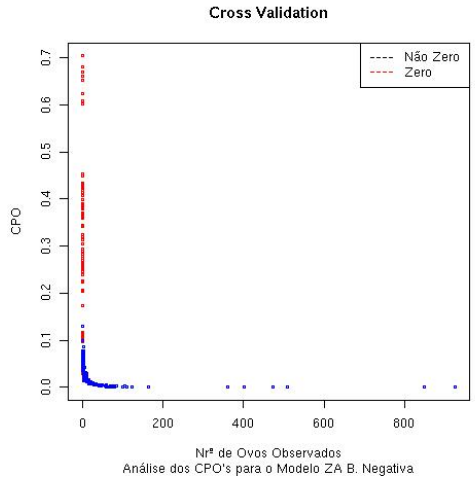


Figura 6.19: CPO ZABN por número de ovos observado por indivíduo

Se compararmos os resultados dos modelos que utilizam a distribuição Binomial Negativa nas Figuras 6.17, 6.18 e 6.19 com as figuras anteriores, relativas aos modelos que utilizam a distribuição de Poisson, verifica-se que nos modelos que utilizam a distribuição Binomial Negativa têm maior capacidade de prever zeros. Porém, nestes modelos registam-se valores baixos das CPO para valores elevados do número de ovos do parasita, tal como nos modelos que utilizam a distribuição Poisson.

Há diferenças interessantes entre os modelos Poisson e Binomial Negativo, se observarmos a "distribuição" das CPO para valores reduzidos do número de ovos. Os modelos com a distribuição Poisson apresentam maior dispersão nos valores estimados do que os modelos que utilizam a Binomial Negativa.

Foram também avaliadas as CPO por indivíduo, tendo sido usada a transformação logarítmica de forma a tornar mais perceptíveis as diferenças entre os modelos. Apresentam-se em paralelo os modelos construídos com a distribuição de Poisson contra os modelos que usam a distribuição Binomial Negativa de forma a facilitar as comparações.

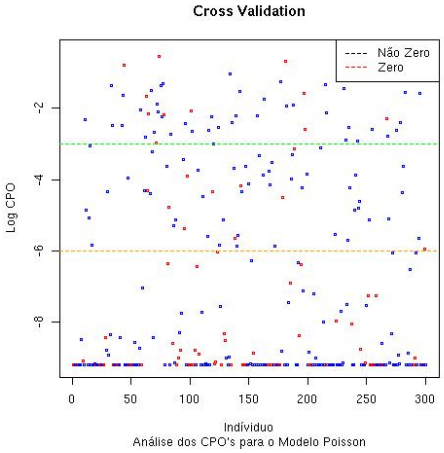


Figura 6.20: Log(CPO) Poisson GLM por Indivíduo

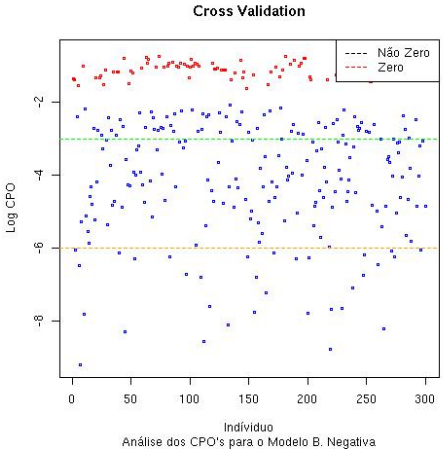


Figura 6.21: Log(CPO) Binomial Negativa GLM por Indivíduo

Nas Figuras 6.20 e 6.21 é visível que o modelo com a distribuição Binomial Negativa mostra ter melhor capacidade preditiva, observando-se um aglomerado das CPO na parte superior do gráfico, o que indica probabilidades mais elevadas. Também se verifica, para os indivíduos sem ovos na amostra de urina, um desempenho superior do modelo que utiliza a distribuição Binomial Negativa.

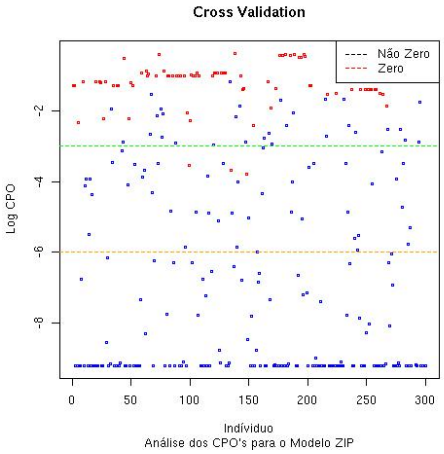


Figura 6.22: Log(CPO) ZIP por Indivíduo

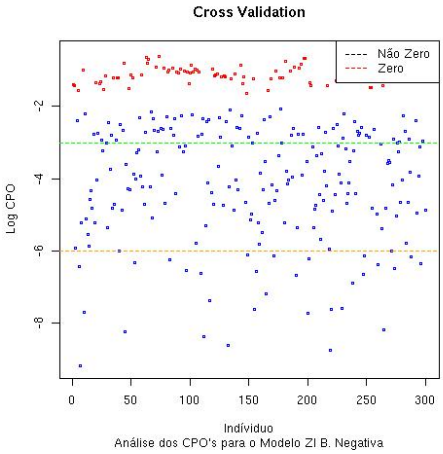


Figura 6.23: Log(CPO) ZIBN por Indivíduo

Nos modelos ZI, tanto para o modelo ZIP como ZIBN, é notória a *nuvem de pontos* associada a *zeros* na parte superior dos gráficos das Figuras 6.22 e 6.23. Estes modelos possuem uma estrutura que modela o aparecimento excessivo de *zeros*, pelo

que indivíduos relativamente aos quais não se observaram ovos nas amostras apresentam valores das CPO mais elevados que os modelos anteriores. O modelo ZIBN tem um desempenho melhor no que respeita ao valores extremos das CPO face ao modelo ZIP. Apesar de melhorias do modelo ZIP face ao Poisson GLM, o modelo ZIP ainda demostra "fragilidade" de previsão com um grupo alargado de valores extremos.

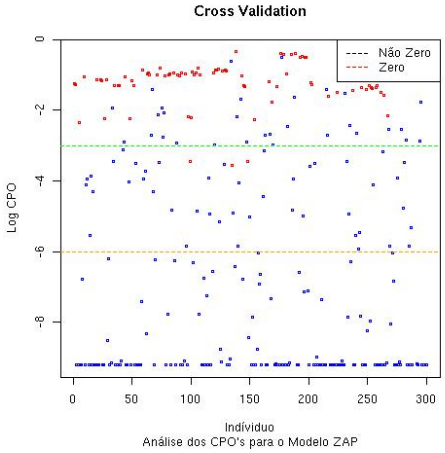


Figura 6.24: Log(CPO) ZAP por Indivíduo

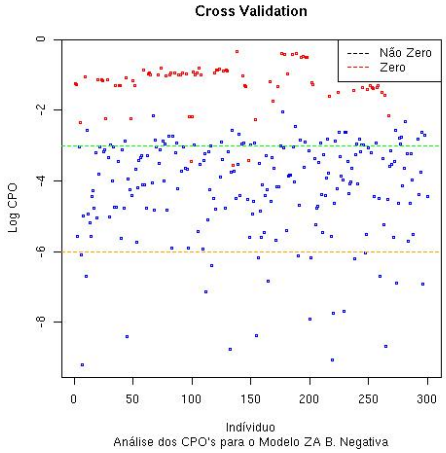


Figura 6.25: Log(CPO) ZABN por Indivíduo

Os modelos ZA demonstram nas Figuras 6.24 e 6.25 uma divisão entre as previsões para os indivíduos com e sem ovos observados nas amostras, o que faz sentido, visto que os modelos ZA estruturalmente fazem essa distinção. Se efectuarmos uma comparação da zona intermédia dos gráficos das Figuras 6.24 e 6.25 é perceptível que o modelo ZABN tem maior densidade de pontos e apresenta menos valores extremos, dando indicações mais favoráveis que o modelo ZAP.

Para dar uma noção quantitativa da *performance* dos modelos, foram estabelecidos arbitrariamente três níveis (baixo, médio e alto) em função dos valores do logaritmo das CPO (log(CPO)) para fazer uma análise da distribuição dos valores estimados. Como é visível nos gráficos anteriores, em que as linhas horizontais nas ordenadas -3 e -6 "partem" o espaço em três fracções, esses mesmos valores foram usados para a análise que se segue.

Nível	Poisson GLM	ZIP	ZAP	Binomial Negativa GLM	ZIBN	ZABN
Baixo]$-\infty$ a -6]	62,7%	47,0%	47,0%	10,7%	10,7%	7,3%
Médio] -6 a -3]	19,3%	15,7%	15,7%	39,7%	39,0%	51,7%
Alto] -3 a 0]	18,0%	37,3%	37,3%	49,7%	50,3%	41,0%

Tabela 6.3: Distribuição por categoria de valores do logaritmo das CPO

Verifica-se que o modelo Poisson GLM tem a maior proporção das *log(CPO)* com valores extremos, o que o caracteriza como o modelo com maiores dificuldades preditivas. Existe semelhança entre os modelos ZIP e ZAP os quais, apesar de serem estruturalmente diferentes, têm o mesmo conjunto de covariáveis, pelo que as diferenças são pequenas. Os modelos que utilizam a distribuição Binomial Negativa apresentam invariavelmente os melhores resultados, com as maiores proporções de valores "Baixos" das *log(CPO)*. Em particular, os modelos Binomial Negativa GLM e ZIBN apresentam o maiores proporções de valores "Altos" e menores de valores "Baixos".

Nível	Poisson GLM	ZIP	ZAP	Binomial Negativa GLM	ZIBN	ZABN
Baixo] $-\infty$ a -6]	72,9%	0,0%	0,0%	0,0%	0,0%	0,0%
Médio] -6 a -3]	11,8%	3,5%	3,5%	0,0%	0,0%	3,5%
Alto] -3 a 0]	15,3%	96,5%	96,5%	100,0%	100,0%	96,5%

Tabela 6.4: Distribuição por categoria de valores do logaritmo das CPO dos Indivíduos sem ovos do parasita observados

Ao analisarmos a Tabela 6.4 relativa apenas aos indivíduos sem *ovos* observados, verifica-se algo bastante semelhante aos resultados globais, mas com maior proximidade do modelos ZIP e ZAP dos modelos com a distribuição Binomial Negativa. Conclui-se que as estruturas dos modelos compostos para modelar o número de *zeros* excessivo são bastante úteis e que a flexibilidade da Binomial Negativa em relação à distribuição de Poisson permite obter modelos mais assertivos.

O produto das CPO é denominado Verosimilhança Pseudo Marginal e é proposta como uma aproximação da distribuição marginal dos dados *p(y)* (Gelfand and Dey, 1994; Gelfand, 1996). Para uma avaliação dos modelos é usado o valor negativo da soma do logaritmo das CPO como forma de comparar a *performance* de modelos. Esta quantidade é denominada *Negative Cross-Validatory Log Likelihood* (NLL) (Nellon. et.

al, 2009). Na tabela seguinte apresenta-se este indicador, calculado para todos os indivíduos (Global) e para os quais não se observou ovos do parasita na amostra (Zeros).

NLL	Poisson GLM	ZIP	ZAP	Binomial Negativa GLM	ZIBN	ZABN
Global	2 026,52	1 596,66	1 593,78	992,52	992,03	1 007,25
Zeros	611,03	103,63	103,49	93,18	93,64	103,61

Tabela 6.5: *Negative cross-validatory Log Likelihood* (NLL)

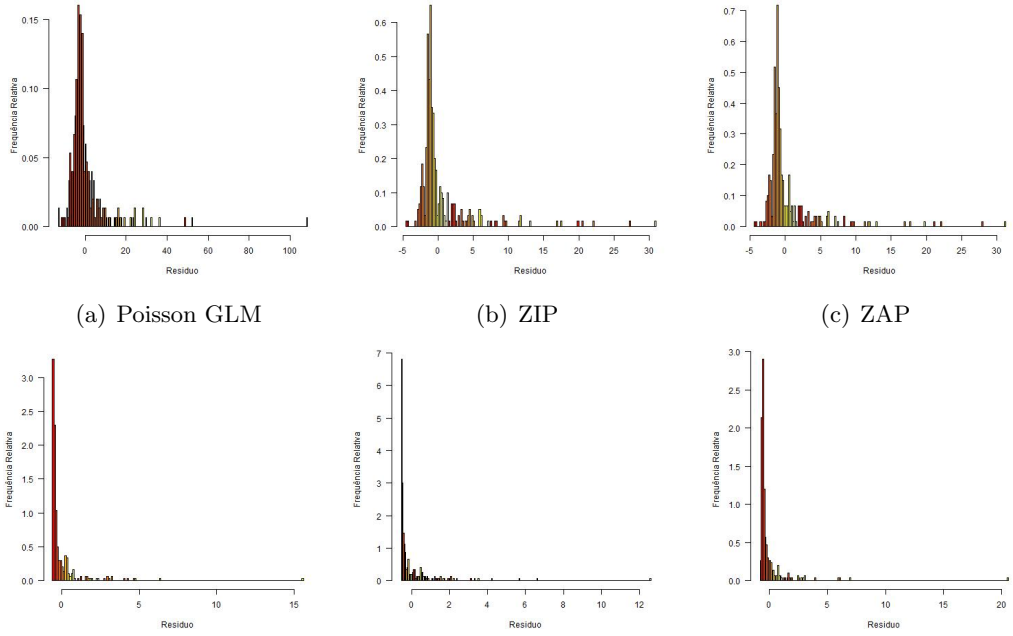
As conclusões anteriores mantêm-se após a análise da Tabela 6.5. De facto, os modelos com a distribuição Binomial Negativa apresentam melhores resultados e o modelo Poisson GLM é o mais ineficiente dos modelos considerados. Através dos valores da NLL também é visível uma grande semelhança entre os modelos Binomial Negativo GLM e o ZIBN e os modelos ZIP e ZAP.

6.2.2 Análise de Resíduos

Nesta subsecção faz-se uma análise dos resíduos padronizados dos diferentes modelos, de forma a avaliar a sua adequabilidade e detectar incoerências. Foram criados gráficos dos resíduos contra os valores esperados e foram analisadas as distribuições dos resíduos padronizados de cada modelo. Os resíduos foram calculados usando a média das *posterioris* das covariáveis como estimativa dos parâmetros.

Em Cameron & Trivedi (2013) é mencionado que para *contagens* os gráficos dos resíduos contra os valores observados não são "informativos", na medida em que há correlação entre os resíduos e os valores observados. Por esta razão pode-se verificar uma tendência ou padrão nos gráficos dos resíduos contra os valores observados, o que nos pode levar a considerar que o modelo apresenta algum "problema", mas trata-se apenas de evidência da mencionada correlação.

6.2.2.1 Distribuição dos Resíduos Padronizados



(a) Poisson GLM (b) ZIP (c) ZAP

(d) Binomial Negativa GLM (e) ZIBN (f) ZABN

Figura 6.26: Histograma dos Resíduos Padronizados

Na Figura 6.26 é possível visualizar que os modelos que utilizam a distribuição de Poisson apresentam maior dispersão nos valores dos resíduos e mais "pontos" com valores elevados que os modelos com a Binomial Negativa. Isto é concordante com o que tem sido verificado nesta análise. Na tabela seguinte é possível comparar os modelos tendo em conta as estatísticas associadas aos resíduos padronizados.

Estatísticas	Poisson GLM	ZIP	ZAP	Binomial Negativa GLM	ZIBN	ZABN
Mediana	-2,1	-0,99	-0,98	-0,44	-0,44	-0,51
Média	0,36	0,43	0,43	-0,02	-0,03	-0,1
Desvio padrão	10,37	4,36	4,37	1,29	1,14	1,54
Máximo	108,3	30,92	31	15,51	12,61	20,68
Minimo	-12,98	-4,53	-4,23	-0,53	-0,53	-0,78
Frequência de Valores < -1.96	158	35	35	0	0	0
Frequência de Valores > 1.96	66	49	50	14	13	12
Soma dos Quadrados dos Resíduos	32311,2	5782,48	5795,04	504,12	397,18	719,51

Tabela 6.6: Análise Estatística dos Resíduos

Na Tabela 6.6 confirma-se o que foi observado na inspecção visual dos resíduos e também nas análises anteriores. Há melhorias notáveis no desempenho dos modelos ZIP e ZAP face ao modelo Poisson GLM devido à introdução das estruturas que modelam o *excesso de zeros*. Os modelos ZIP e ZAP têm resíduos com uma média mais próxima de zero e com menor variabilidade que o modelo Poisson GLM.

6.2.2.2 Comparação dos Resíduos com os Valores Esperados

A distribuição dos valores esperados tem uma cauda direita pesada, podendo os valores elevados sugerir visualmente um padrão inexistente. De forma a reduzir este efeito foram também construídos gráficos dos resíduos padronizados contra o logaritmo dos valores esperados.

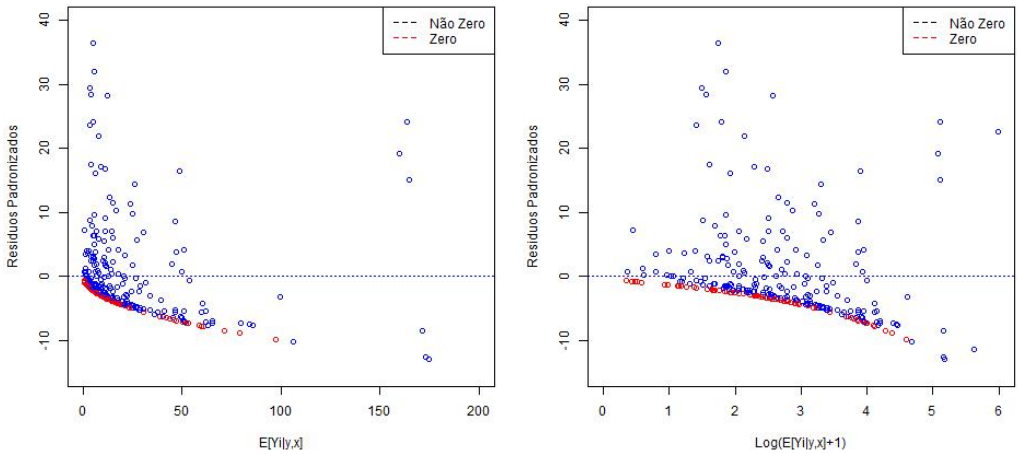


Figura 6.27: Resíduos Padronizados Poisson GLM *contra* Valores Esperados

Na Figura 6.27, relativa ao modelo Poisson GLM, observa-se um padrão à medida que os valores esperados aumentam, o que pode ser interpretado como variabilidade não explicada pelo modelo (heterocedasticidade). Esta tendência negativa é indicativa de que a variância cresce mais rapidamente que a média. Isto é particularmente visível para os indivíduos em que não se observaram ovos na amostra (a vermelho no gráfico), o que mostra a falta de capacidade do modelo Poisson para "lidar" com o *excesso de zeros*.

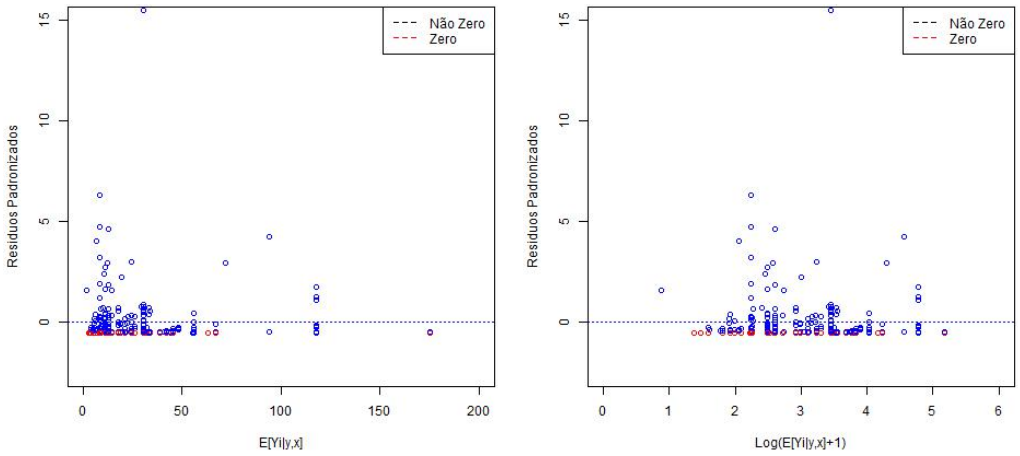


Figura 6.28: Resíduos Padronizados Binomial Negativa GLM *contra* Valores Esperados

No modelo Binomial Negativo GLM, cujos resíduos podem ser observados na Figura 6.28, a tendência é quase inexistente. Porém, é possível visualizar um conjunto de resíduos elevados, o que mostra algumas limitações na capacidade preditiva do modelo, algo que também foi verificado na análise das CPO.

Na comparação entre os gráficos do logaritmo dos valores esperados contra os resíduos é notória a melhoria do modelo Binomial Negativo GLM em relação ao modelo Poisson GLM, o que se deve à sua maior flexibilidade, ajustando-se claramente melhor às características dos dados.

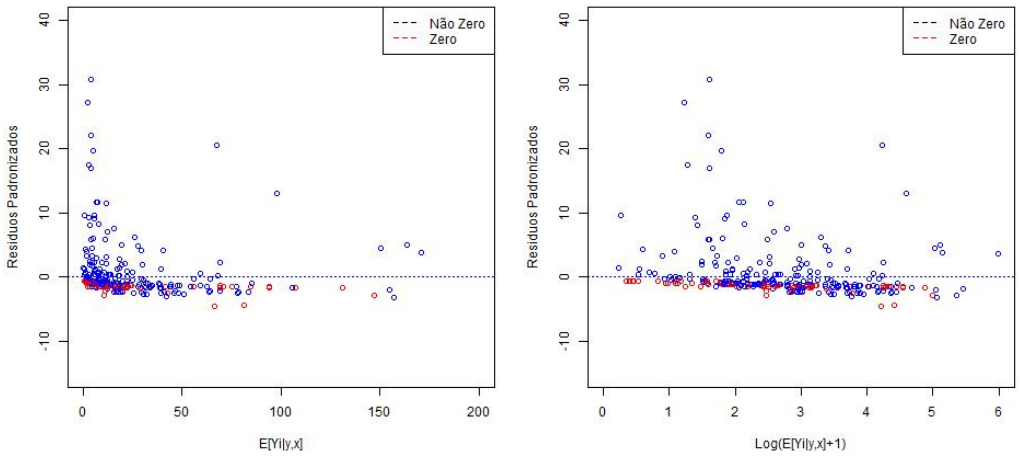


Figura 6.29: Análise Resíduos Padronizados ZIP *contra* Valores Esperados

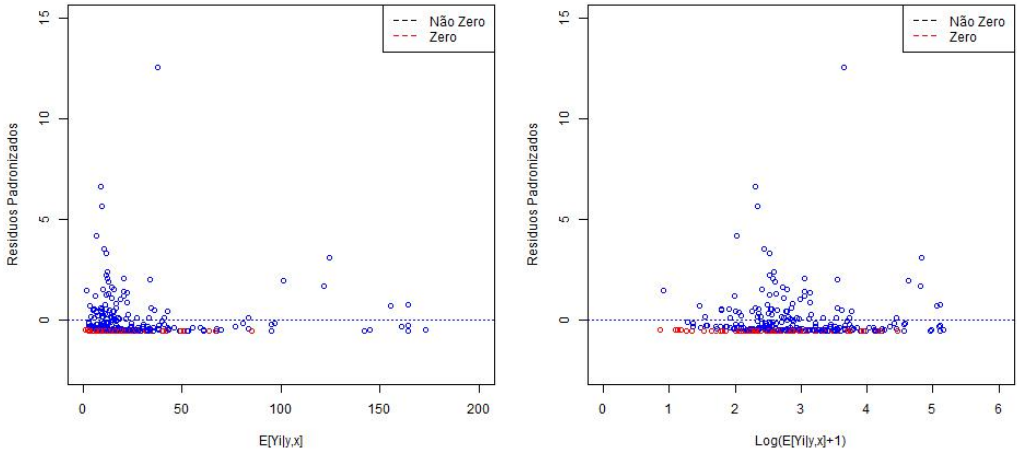


Figura 6.30: Análise Resíduos Padronizados ZIBN *contra* Valores Esperados

O padrão dos resíduos em função do aumento dos valores observados do modelo ZIP é muito menos evidente que no modelo Poisson GLM (comparando as Figuras 6.27 e 6.29). A introdução de uma estrutura que modela o *excesso de zeros* permite em certa medida corrigir o padrão que indica dificuldades em explicar a variabilidade nos dados. Ainda assim, verifica-se no modelo ZIP uma ligeira tendência de redução do valor dos resíduos com o aumento do valor esperado. Esta tendência não é perceptível no modelo ZIBN como pode ser visualizado na Figura 6.30.

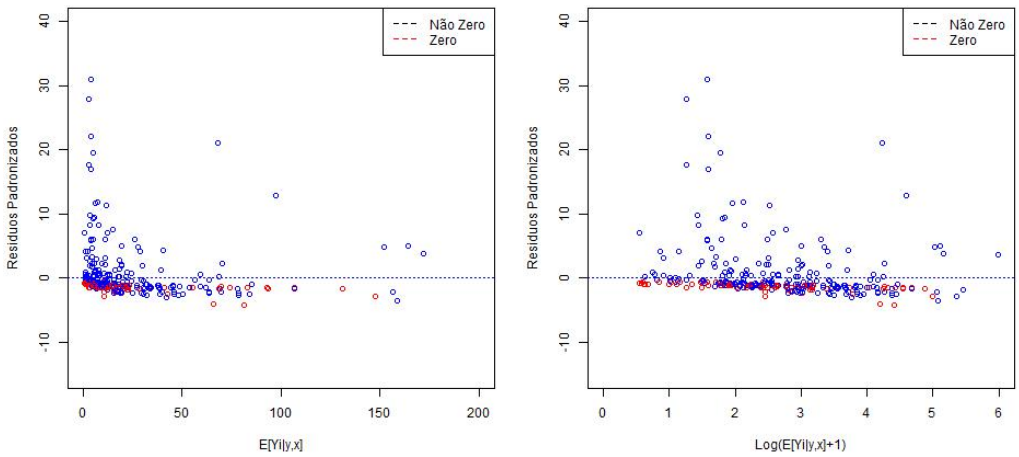


Figura 6.31: Análise Resíduos Padronizados ZAP *contra* Valores Esperados

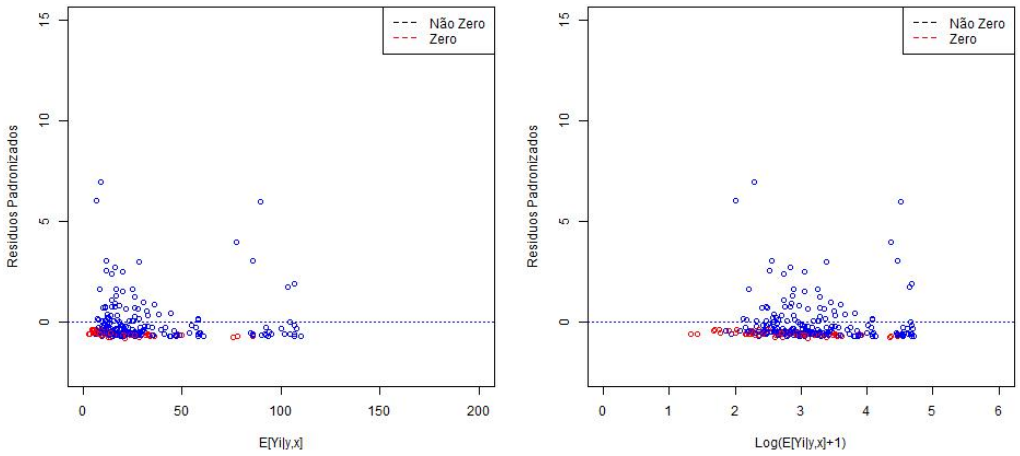


Figura 6.32: Análise Resíduos Padronizados ZABN *contra* Valores Esperados

Os modelos ZA conduzem a resultados muito semelhantes dos modelos ZI. Na representação gráfica dos resíduos do modelo ZABN não é perceptível existir um padrão que indicia problemas estruturais com o modelo, mas é visível a existência de resíduos com valores elevados. Nos resíduos do modelo ZAP na Figura 6.31, há uma ligeira tendência de redução dos valores dos resíduos padronizados com o aumento dos valores esperados. Esta tendência é indicativa de problemas de adequabilidade do modelo, sendo justificável pela rigidez do modelo Poisson, no que respeita à modelação da variabilidade, algo muito semelhante ao que observa no modelo ZIP.

Em resumo, os modelos que recorrem à distribuição de Poisson, apresentam problemas em modelar as características dos dados. É notável a melhoria do desempenho destes modelos quando estão associados a uma estrutura que permite modelar o *excesso de zeros*. Ainda assim, são visíveis deficiências, havendo nos gráficos dos resíduos sinais de heterogeneidade não controlada pelos modelos ZIP e ZAP. Os modelos que recorrem à distribuição Binomial Negativa têm um comportamento superior, não se verificando na análise dos resíduos sinais de incoerência dos modelos. São modelos mais flexíveis, sendo superiores para analisar os dados, tal como se tem verificado ao longo deste trabalho.

6.2.3 Binomial Negativa Sobre Parametrizada

Foi implementada mais uma alternativa para modelar a carga parasitária de forma a tentar melhorar os resultados obtidos. Baseia-se no modelo Binomial Negativo GLM, dado ser o modelo mais parcimonioso e ter demonstrado bons resultados relativamente aos outros modelos. Foi adicionada uma componente que permite modelar o parâmetro de dispersão em função das covariáveis. Assim, este modelo permite modelar diferentes níveis de dispersão, o que torna o modelo Binomial Negativo ainda mais flexível.

Como o parâmetro de dispersão tem de ser positivo foi utilizada a função de ligação *log-linear*. O modelo de regressão é muito semelhante ao Binomial Negativo GLM com esta componente adicional. Ficamos com o seguinte modelo semelhante ao apresentado em (4.7) para o Binomial Negativo GLM:

$$\begin{aligned} &Y_i \sim Neg.Binomial\left(\tfrac{k_i}{k_i+\mu_i}, k_i\right) \\ &\quad com \qquad log(\mu_i) = \beta_0 + \textstyle\sum_{j=1}^{m} \beta_j x_{ij} \qquad e \qquad log(k_i) = \alpha_0 + \textstyle\sum_{j=1}^{m} \alpha_j x_{ij} \end{aligned} \tag{6.1}$$

Foi usada apenas uma cadeia, devido à necessidade de inicializar o parâmetro de dispersão r_i como um inteiro positivo, o que devido à função de ligação considerada e à complexidade do preditor linear torna difícil de conseguir diferentes pontos para inicializar as cadeias. Foi gerada uma cadeia com um *burn-in* longo e obtida uma amostra também de "elevada" dimensão. Para cada coeficiente do preditor linear foi usada como *priori* a distribuição Normal, centrada em zero e com variância elevada ($\mathcal{N}(0, 1/0.0001)$), tal como para os restantes modelos.

Seguidamente apresenta-se a tabela com os valores médios das *posterioris* do modelo obtido. A tabela está dividida em duas colunas com os parâmetros que modelam os parâmetros de dispersão e os que modelam a componente do valor médio de ovos.

Este modelo apresenta melhorias face ao modelo Binomial Negativa GLM, em termos de AIC (2018,2) e DIC (2017,7) como se pode verificar na Tabela 6.1, mas à custa de mais 15 parâmetros, pelo que o BIC deste último modelo é mais elevado que no modelo BN GLM (BIC 2140,5).

Covariáveis	Binomial Negativa Sobre Parametrizada	
	Dispersão	Média
Ordenada na origem	-0,8423	3,94
Género(F)	a)	a)
Género (M)	-0,3561	0,06391
Idade	0,003657	-0,01124
Água Canal. (Sim)	a)	a)
Água Canal. (Não)	—	—
WC (Dentro)	a)	a)
WC (Fora)	-0,7729	0,7238
Cont. com água (Rio)	a)	a)
Cont. com água (Lagoa)	-0,6939	-0,8041
Cont. com água (Tanque)	0,1062	-1,553
Conhec. da doença (Sim)	a)	a)
Conhec. da doença (Não)	—	1,098
Hematúria (Negativo)	a)	a)
Hematúria (Positivo)	—	1,294
Prof. (Func. Público)	a)	a)
Prof. (Estudante)	—	-1,566
Prof. (Agricultor)	—	-1,95
Prof. (Trab. doméstico)	—	-1,552
Prof. (Outras)	—	-0,8782
Motivo (Buscar água)	a)	a)
Motivo(Higiene Pessoal)	1,098	-1,342
Motivo (Lavar roupa)	1,294	-0,8713
Motivo (Pescar)	-1,95	-1,277
Motivo (Nadar)	-1,552	-0,7065
Motivo (Todos os anteriores)	-1,566	-1,331
Província (Luanda)	a)	a)
Província (Bengo)	—	—
Província (Kwanza Sul)	—	—
Naturalidade (Luanda Bengo)	a)	a)
Naturalidade (Bié Huambo Moxico)	—	—
Naturalidade (Norte)	—	—
Naturalidade (Sul)	—	—
$\overline{D(\bar{\theta})}$	1950,61	
$\overline{D(\theta)}$	1978	
DIC	2005,38	
AIC	2006,61	
BIC	2270,02	
Nr^o Parâmetros	28	
burn in	30000	
Tamanho Amostra	6000	
Thinning	275	
Nr^o Cadeias	1	

[a)] Categoria de Referência
— Nível da covariável não utilizado

Tabela 6.7: Valor médio da *posteriori* dos parâmetros do modelo Binomial Negativo Sobre Parametrizado

6.2.3.1 CPO Binomial Negativa Sobre Parametrizada

O comportamento das CPO é semelhante à dos modelos anteriormente analisados. Porém regista-se alguma melhoria face aos resultados anteriores.

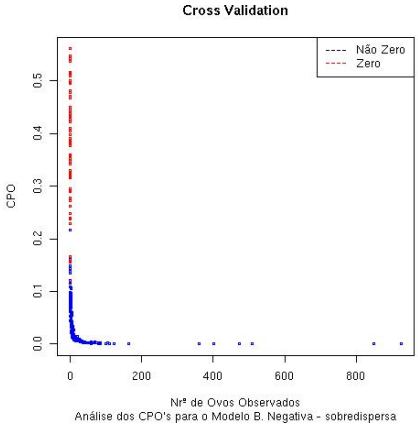


Figura 6.33: CPO Binomial Negativa Sobre Parametrizada por Número de ovos observado

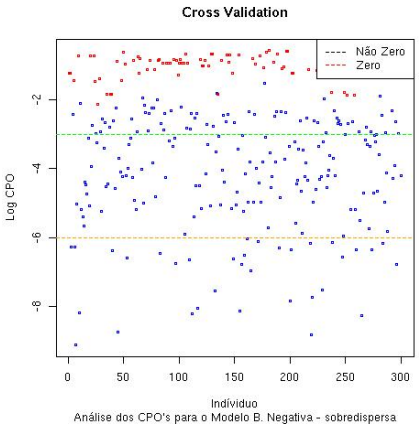


Figura 6.34: Log(CPO) Binomial Negativa Sobre Parametrizada por Indivíduo

É visível a diminuição dos valores das CPO à medida que o número de ovos observado aumenta e também a existência de um número substancial de elementos com valores das *log(CPO)* negativos. No entanto, o comportamento deste modelo é melhor que o dos anteriores. Há melhorias no indicador NLL, sendo mais reduzido do que o obtido para qualquer um dos modelos analisados anteriormente. Sendo o menor valor registado anteriormente de 992,52 para o modelo Binomial Negativo GLM.

Nível	Todas a Observações	Apenas Zeros
Baixo] $-\infty$ a -6]	11,3%	0%
Médio] -6 a -3]	38,3%	0%
Alto] -3 a 0]	50,3%	100,0%
NLL	987,71	88,94

Tabela 6.8: Distribuição por categoria de valores do logaritmo das CPO do modelo Binomial Negativo Sobre Parametrizado e Estimativa NLL

6.2.3.2 Resíduos Binomial Negativa Sobre Parametrizada

Na análise dos resíduos deste modelo também se verificam melhorias face aos modelos anteriores, sendo que a Binomial Negativa Sobre Parametrizada apresenta melhores resultados. Ainda assim, observa-se a presença de resíduos positivos elevados, o que indica que existem casos em que o modelo subestima o número de ovos do parasita.

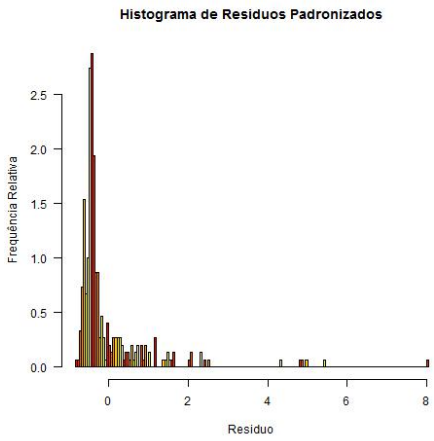


Figura 6.35: Histograma dos resíduos Binomial Negativa Sobre Parametrizada

Estatísticas	Binomial Negativa Sobre Parametrizada
Mediana	-0,39
Média	-0,04
Desvio padrão	1
Máximo	8,01
Minimo	-0,84
Valores < -1.96	0
Valores > 1.96	13
Soma dos quadrados dos resíduos	305,72

Tabela 6.9: Estatística descritiva dos Resíduos Modelo Binomial Negativo Sobre Parametrizado

As estatísticas dos resíduos indicam também um desempenho superior deste modelo, em particular quando analisamos a soma dos quadrados dos resíduos, que se deve ao controlo da dispersão que este modelo permite obter. Na Figura 6.37, respeitante à representação dos resíduos padronizados contra o logaritmo dos valores esperados, não se verifica um padrão indicativo de incoerência no modelo, tal como se observou nos modelos que utilizam a distribuição Binomial Negativa.

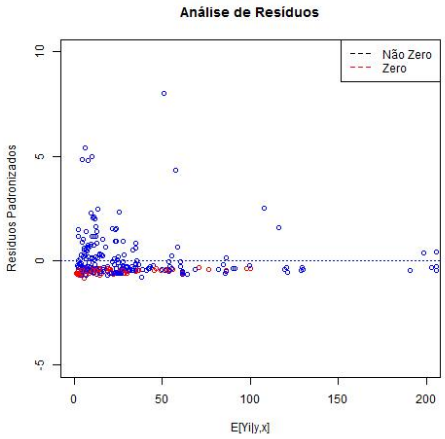


Figura 6.36: Resíduos Binomial Negativa Sobre Parametrizada *contra* Valores Esperados

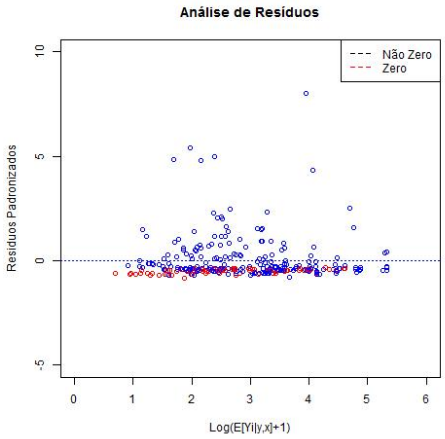


Figura 6.37: Resíduos Binomial Negativa Sobre Parametrizada por Logaritmo *contra* Valores Esperados

Nesta secção apresentou-se mais um modelo alternativo aos inicialmente perspectivados para este trabalho. Esta alteração do modelo Binomial Negativo GLM é relativamente fácil de implementar devido à possibilidade de adicionar mais um preditor linear nas condições de (6.1) no programa BUGS. Apesar da melhoria dos resultados, com este modelo foi aumentada a complexidade o que não serve o princípio de parcimónia e também é dificultada a interpretação do modelo.

7

Discussão

Conseguiu-se implementar com sucesso em BUGS e R um conjunto de modelos com características complexas de forma a modelar dados de natureza epidemiológica. Verificou-se ao longo desta análise que o modelo Poisson, usualmente utilizado para modelar variáveis de contagem, não tem capacidade para dar resposta às características particulares dos dados sobre a *schistosomose* e que as alternativas com recurso à distribuição Binomial Negativa forneceram os melhores resultados.

A introdução das estruturas para modelar o *excesso de zeros* permitiu melhorar de forma marcada a *performance* do modelo Poisson GLM. Verificou-se que os modelos ZIP e ZAP apresentaram resultados muito semelhantes, quer em termos de indicadores como AIC, DIC, BIC (Tabela 6.1), como pelo indicador NLL (Tabela 6.5), sendo os resultados para as observações respeitantes aos *zeros* muito semelhantes. Isto deve-se à incapacidade do modelo Poisson para lidar com o *excesso de zeros*, por este motivo as estruturas do modelo logístico nos modelos ZIP e ZAP estão encarregues de explicar o aparecimento de *zeros*. Assim o modelo ZIP ficou "dividido" em duas partes, daí os resultados serem tão semelhantes ao modelo ZAP.

Na análise de resíduos dos modelos que utilizam a distribuição Poisson, verificou-se na representação dos resíduos padronizados *contra* os valores observados, uma tendência negativa, o que é um sinal de que os modelos não conseguem modelar convenientemente a variabilidade dos dados. Há melhoria nos modelos ZIP e ZAP em relação ao modelo Poisson GLM, sendo ainda assim, notória a dificuldade em modelar a variabilidade,

algo que é visível nas Figuras 6.29 e 6.31 respeitantes aos resíduos dos modelos ZIP e ZAP, respectivamente.

Os modelos baseados na distribuição Binomial Negativa apresentam invariavelmente os melhores resultados em todos os indicadores, com menos parâmetros e sem sinais evidentes de incoerência no que respeita à análise dos resíduos. Nota-se uma semelhança entre os modelos Binomial Negativo GLM e o modelo ZIBN, o que é compreensível visto que a estrutura do modelo logístico que modela o *excesso de zeros* tem um peso negligenciável no modelo ZIBN. Na Tabela 6.5, no que respeita ao valor NLL calculado para os *zeros*, não há diferenças substanciais entre o modelo ZIBN e o Binomial Negativo GLM, isto apesar do modelo ZIBN ter uma estrutura adicional para modelar os *zeros*. Este facto pode ser visto como indicação de que o modelo Binomial Negativo GLM é capaz de lidar com o *excesso de zeros* nestes dados, algo que também é observável na representação gráfica das CPO em 6.21, com uma *nuvem de pontos* na parte superior da figura muito semelhante aos modelos ZI e ZA.

O modelo ZABN supera o desempenho dos modelos ZIP e ZAP devido à utilização da distribuição Binomial Negativa para modelar o número de ovos observados. São de notar semelhanças nas estimativas para os modelos ZIP, ZAP e ZABN para a componente do modelo logístico apresentadas na Tabela 6.2 e da performance destes modelos para explicar os *zeros*. Algo que é confirmado na análise das CPO na Tabela 6.4 e do indicador NLL na Tabela 6.5.

Implementou-se mais uma alternativa aos modelos inicialmente previstos, a Binomial Negativa Sobre Parametrizada. Neste modelo foi construída uma estrutura que permite modelar o parâmetro de dispersão da distribuição Binomial Negativa em função da covariáveis. Em termos dos indicadores como AIC ou DIC obteve-se algumas melhorias face aos restantes modelos. No entanto é uma alternativa complexa com um conjunto de parâmetros elevado face aos outros modelos. Ficou visível como em BUGS é simples adicionar esta componente e obter as estimativas das distribuições *a posteriori* dos parâmetros adicionais deste modelo.

Outra alternativa para a análise de *contagens* em que se observa sobredispersão é o modelo Poisson Generalizado, no qual existe uma componente adicional no modelo Poisson, que permite modelar a variância de forma independente da média. A Poisson Generalizada tem função de massa de probabilidade da seguinte forma:

$$P(Y=y)=\left(\frac{\mu}{1+\alpha\mu}\right)^{y}\frac{(1+\alpha y)^{y-1}}{y!}e^{\frac{-\mu(1+\alpha y)}{1+\alpha\mu}}, \qquad y=0,1,2,...$$

Este modelo tem $E(Y)=\mu$ e $Var(Y)=Var(\mu(1+\alpha\mu)^2)$, com α denominado como parâmetro de dispersão. É simples de verificar que se $\alpha=0$ este modelo reduz-se o modelo Poisson, se $\alpha>0$ a variância é maior que o valor médio e se $\alpha<0$ verifica-se o oposto, este modelo pode acomodar tanto sobredispersão como subdispersão.

A Poisson Generalizada foi implementada com todas as covariáveis. No entanto, não se conseguiu prosseguir com a análise devido a questões técnicas ao nível de programação e à falta de tempo para resolver esses entraves. O programa criado para este modelo e os resultados preliminares serão disponibilizados nos anexos em suporte digital. Trata-se de uma alternativa viável para uma análise futura, inclusive a combinação deste modelo com um modelo logístico para modelar o *excesso de zeros* poderá ser uma possibilidade a ter em conta.

A escolha sobre que modelo será superior recai sobre o modelo Binomial Negativo GLM. Este é considerado superior devido à sua simplicidade e capacidade de se ajustar às características particulares dos dados em análise. É de referir que se obteve resultados semelhantes ao modelo ZIBN, que possui uma estrutura específica para modelar o excesso de *zeros*. Deste modo é também de questionar se as covariáveis que foram usadas para modelar o *excesso de zeros* serão as mais indicadas. Questões como a persistência de sintomas da *schistosomose* ou o tempo decorrido desde o aparecimento dos primeiros sintomas da doença poderão dar mais informação no sentido de explicar o *excesso de zeros*. Podendo providenciar uma perspectiva do espaço de tempo desde a infecção e o respectivo desenvolvimento e incubação do parasita dentro do corpo do hospedeiro definitivo até ao inicio da excreção de ovos de schistosoma.

Bibliografia

[1] Bolstad, W. (2004) Introduction to Bayesian Statistics. John Wiley & Sons, Inc. *Cap. 4.*

[2] Bruun, B., Aagaard-Hansen, J. (2008). The social context of schistosomiasis and its control - *An introduction and annotated bibliography* -WHO Library Cataloguing-in-Publication Data. pp.1-10 e 33-41. Obtido em http://www.who.int/tdr/publications/documents/social-context-schistosomiasis.pdf, em 31 de Maio de 2013.

[3] Cardoso, S. (2010). Schistosomose urinária e helmintoses intestinais: contribuição para o estudo clínico-epidemiológico e da respostaimune humoral na comunidade angolana. *IHMT: HMM - Dissertações de Mestrado. Instituto de Higiene e Medicina Tropical. Universidade Nova de Lisboa.*

[4] CDC - Center for Disease Control and Prevention Website. *http://www.cdc.gov/parasites/schistosomiasis/*

[5] Chitsulo, L., Engels, D., Montresor, A. and Savioli, L. (2000) The global status of schistosomiasis and its control. Acta Tropical, 77: 41-51. `http://www.sciencedirect.com/science/article/pii/S0001706X00001224`

[6] Cameron, C. and Trivedi, P. (2013). Regression Analysis of Count Data, 2nd edition, Cambridge University Press. *pp. 178 a 188.*

[7] Congdon, P. (2003). Applied Bayesian Modelling. *Wiley series in Probability and Statistics*, *Cap. 1 e 3.*

[8] Congdon, P. (2005). Bayesian Models for Categorical Data, *Wiley series in Probability and Statistics*, *Cap. 5.*

[9] Cook, J. (2009). Notes on the Negative Binomial Distribution. Informação obtida a 1 de Janeiro de 2013 da World Wide Web: www.johndcook.comnegative_binomial.pdf .

[10] McManus, D. and Loukas A. (2008). Current Status of Vaccines for Schistosomiasis - Clin. Microbiol. Rev. 2008 January; 21(1): 225–242. American Society for Microbiology (ASM). Acedido na web em 2 de julho de 2013: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2223839/?report=reader*.

[11] Efron, B. (1986). Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association, 81, 709–72.

[12] Exemplos de código e aplicações em Bugs *http://www.mrc-bsu.cam.ac.uk/bugs/*.

[13] Figueiredo, Jacinta T. (2008). Contribuição para o estudo da epidemiologia de morbilidade da Schistosomose vesical na população adulta de Angola. Províncias de Luanda, Bengo e Kwanza Sul. Instituto de Higiene e Medicina Tropical. Universidade Nova de Lisboa.

[14] Frees, E. (2009). Regression Modeling With Actuarial and Financial Applications; Cambridge University Press. Cap 12 Count Dependent Variables, Overdispersion and Negative Binomial Models, *pp.352-352*.

[15] Gelman, A. , Carlin, J. , Stern, H. , Rubin, D. (2003). Bayesian Data Analysis, Chapman & Hall CRC Texts in Statistical Science. *Cap. 1 e 11* .

[16] Website da WHO sobre Schistosoma (Maio de 2013). http://www.who.int/topics/schistosomiasis/en/

[17] Santos J., Chaves J. , Videira, M., et al (2012)- Schistosomose Haematobium e carcinoma da bexiga. Acta Urológica – Março de 2012 – 4: 13–17.

[18] King, C. (2010). Parasites and poverty: The case of Schistosomiasis - Acta Trop. 2010 February; 113(2): 95–104.

[19] Krush, J. (2010). Doing Bayesian Data Analysis. Elsevier Inc. *Cap. 23*.

[20] Mostafa, M., Sheweita S., O Connor, P. (1999). Relationship between Schistosomiasis and Bladder Cancer, Clinical Microbiology Review January 1999 vol. 12 no. 1 97-111 - American Society of Microbiology. Acedido na web em *http://cmr.asm.org* - 15 junho 2013

[21] Mahmoud, A. (2001) Schistosomiasis. Tropical Medicine Science and Practice: Imperial College Press, *Cap. 1, 2 e 4.*

[22] Manual de Open Bugs: *http://www.openbugs.info/Manuals/Manual.html*

[23] Plummer, M., Best, N. , Cowles, K., Vines, K., Sarkar, D., Almond, R. , Output analysis and diagnostics for MCMC - Informação obtida a 1 de Março de 2013 da da World Wide Web: http://cran.r-project.org/web/packages/coda/coda.pdf .

[24] Muller, D. (2007). Processos Estocásticos e Aplicações. II Série, Nr^{o}3 Colecções Económicas, Almedina, *pp.101-102.*

[25] Neelon, B., O'Malley A., Normand, Sharon-Lise T(2009). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling.*

[26] Ntzoufras, I. (2009) Bayesian Modeling Using WinBUGS. John Wiley & Sons, Inc.

[27] Olivença, D. (2011). Modelos com Excesso de Zeros e modelos de Duas Partes - a sua utilização no estudo da Schistosomose. *Dissertação de Mestrado. FCUL DEIO. Universidade Lisboa.*

[28] Paulino, C. , Turkman M., Murteira, B. (2003). Estatística Bayesiana. Fundação Calouste Gulbenkian.

[29] Neal, R. (2003). Slice Sampling. *The Annals of Statistics.* Vol. 31, No. 3, 705–767.

[30] Rowe, D. (2003). Multivariate Bayesian Statistics. Chapman & Hall CRC. *Cap. 4 e 8.*

[31] Shiff, C., Veltri, R., Naples, J., Quartey, J., Otchere, J., Anyan, W., Marlow, C., Wiredu, E., Adjei, A., Brakohiapa, E., Bosompem, K. - Ultrasound verification of bladder damage is associated with known biomarkers of bladder cancer in adults

chronically infected with Schistosoma haematobium in Ghana. Tropical Medicine and Hygiene, 100 (9): 847-854.

[32] Spiegelhalter, D. , Best, N., Carlin B., Van Der Linde A. (2002). Bayesian measures of model complexity and fit, J.R. Statistical Society B 64, Part 4, *pp. 583-639.*

[33] Torgo, L. (2009). A Linguagem R. Programação para análise de Dados. Escolar Editora.

[34] Turkman, M. (2000) Modelos Lineares Generalizados - da teoria à prática; . *VIII Congresso anual de Estatística.*

[35] Velez, A. (2010). Bioecologia e caracterização molecular de Bulinus globosus de Angola. Lisboa. Instituto de Higiene e Medicina Tropical.

[36] WER (2013), WHO - Weekly epidemiological record, No. 8, Fevereiro de 2013, 88, *pp. 81–88 http://www.who.int/wer/en/* - obtido em 1 de Junho de 2013.

[37] WHO Website `http://www.who.int/schistosomiasis/epidemiology/en/`

[38] WHO, Expert Committee on the Control of Schistosomiasis (2002). Prevention and control of schistosomiasis and soil-transmitted helminthiasis: report of a WHO expert committee. Geneva, Switzerland, World Health Organization.

[39] Zuur, A. (2009). Mixed Effects Models and Extension in Ecology with R. Statistics for Biology and Health.*pp. 261-293*

8

Anexos

8.1 Comparação Abordagem Clássica e Bayesiana

Covariáveis	Poisson GLM		Bin. Negativa GLM		ZIP				ZIBN				ZAP				ZABN			
					Zeros		Média		Zeros		Média		Zeros		Média		Zeros		Média	
	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana	Clássica	Bayesiana
Ordenada na origem	1,651	1,927	3,581	3,183	-0,95	-0,9151	1,333	1,772	-301,97	-238,1	2,602	2,652	0,965	-0,9776	1,328	1,765	0,965	-0,982	0,568	2,675
Género(F)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Género (M)	0,458	0,7236	-0,022	0,3988	0,162	0,1554	0,569	0,8784	-65,752	-29,54	0,146	0,406	-0,2	0,2074	0,573	0,882	-0,2	0,209	-0,406	-0,198
Idade	-0,013	-0,01927	-0,013	-0,01713	0,03	0,0007145	-0,006	-0,0144	4,897	-6,471	-0,005	-0,017	-0,005	0,00465	-0,006	-0,01442	-0,005	0,004	-0,029	-0,021
Água Canal. (Sim)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Água Canal. (Não)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
WC (Dentro)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
WC (Fora)	0,776	0,8237	—	—	—	—	0,92	0,9469	—	—	0,67	0,527	—	—	0,884	0,9494	—	—	0,746	0,711
Cont. com água (Rio)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Cont. com água (Lagoa)	-0,402	-0,3653	-0,651	-0,5597	1,744	1,837	-0,102	0,1633	—	—	-0,367	-0,467	-1,801	1,881	-0,108	0,1647	-1,801	1,872	—	—
Cont. com água (Tanque)	-1,259	-1,525	-0,983	-1,292	0,358	0,3741	-1,196	-1,628	—	—	-1,193	-1,263	-0.368	0,3698	-1,199	-1,63	-0,368	0,369	—	—
Conhec. da doença (Sim)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Conhec. da doença (Não)	1,218	1,142	–	0,7588	—	—	1,258	1,269	—	—	1,112	0,887	—	—	1,36	1,273	—	—	—	0,634
Hematúria (Negativo)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Hematúria (Positivo)	1,322	1,323	1,296	1,349	-1,163	-1,233	1,183	1,116	—	—	1,276	1,296	1,213	-1,295	1,184	1,118	1,213	-1,292	1,283	1
Prof. (Func. Público)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Prof. (Estudante)	0,106	0,1468	—	—	—	—	0,335	0,3624	—	—	-0,811	—	—	—	0,338	0,3631	—	—	—	—
Prof. (Agricultor)	-0,441	-0,4289	—	—	—	—	-0,224	-0,2351	—	—	-1,072	—	—	—	-0,221	-0,235	—	—	—	—
Prof. (Trab. doméstico)	0,24	0,3825	—	—	—	—	0,607	0,6893	—	—	-0,609	—	—	—	0,612	0,6912	—	—	—	—
Prof. (Outras)	-0,139	-0,09586	—	—	—	—	0,186	0,2561	—	—	-0,804	—	—	—	0,194	0,2575	—	—	—	—
Motivo (Buscar água)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Motivo(Higiene Pessoal)	-0,3	-0,6978	-0,442	-0,8909	—	—	-0,456	-0,9026	—	—	-0,582	-0,888	—	—	-0,46	-0,9048	—	—	—	—
Motivo (Lavar roupa)	-0,878	-1,45	-0,662	-0,7472	—	—	-1,222	-1,755	—	—	-0,342	-0,653	—	—	-1,227	-1,761	—	—	—	—
Motivo (Pescar)	-1,263	-1,974	-0,742	-1,421	—	—	-1,448	-2,246	—	—	-0,898	-1,3991	—	—	-1,457	-2,255	—	—	—	—
Motivo (Nadar)	0,053	-0,4404	0,342	-0,6223	—	—	0,349	-0,2587	—	—	0,132	-0,682	—	—	0,345	-0,2588	—	—	—	—
Motivo (Todos os anteriores)	—	-1,633	—	-1,667	—	—	—	-1,956	—	—	—	-1,615	—	—	—	-1,963	—	—	—	—
Província (Luanda)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Província (Bengo)	0,223	0,08949	—	—	-0,377	-0,3143	0,012	-0,1127	—	—	—	—	0,391	-0,4125	-0,01	-0,1145	0,391	-0,406	—	—
Província (Kwanza Sul)	0,454	0,351	—	—	-1,321	-1,46	0,29	0,1661	—	—	—	—	1,324	-1,407	-0,287	0,167	1,324	-1,401	—	—
Naturalidade (Luanda Bengo)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)	a)
Naturalidade (Bié Huambo Moxico)	0,311	0,3545	—	—	—	—	0,528	0,6097	—	—	—	—	—	—	0,53	0,6115	—	—	—	—
Naturalidade (Norte)	-0,477	-0,496	—	—	—	—	-0,612	-0,6078	—	—	—	—	—	—	-0,615	-0,6091	—	—	—	—
Naturalidade (Sul)	0,463	0,5401	—	—	—	—	0,183	0,3138	—	—	—	—	—	—	0,183	0,3119	—	—	—	—
Parâmetro de dispersão: r	—	—	—	0,2823	—		—	0,2823	—	—	—	0,285	—	—	—	—	—	—	—	0,508

[a)] Valor Base

— Covariável não utilizada

Tabela 8.1: Valores dos parâmetros dos modelos Clássico *contra* Bayesiano

8.2 Estatísticas das *Posterioris* dos Modelos

8.2.1 Estatísticas Poisson GLM

	Poisson GLM					
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	1,927	0,094	0,00087	1,743	1,928	2,11
Género(F)	a)	-	-	-	-	-
Género (M)	0,723	0,04	0,000345	0,644	0,723	0,804
Idade	-0,019	0,001	0,000012	-0,022	-0,019	-0,016
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	-	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	0,823	0,046	0,000389	0,732	0,824	0,914
Contacto com água (Rio)	a)	-	-	-	-	-
Contacto com água (Lagoa)	-0,365	0,056	0,000488	-0,477	-0,365	-0,252
Contacto com água (Tanque)	-1,525	0,045	0,000393	-1,614	-1,526	-1,436
Conhecimento da doença (Sim)	a)	-	-	-	-	-
Conhecimento da doença (Não)	1,142	0,043	0,000374	1,058	1,142	1,227
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	1,323	0,029	0,000269	1,265	1,323	1,38
Profissão (Func. Público)	a)	-	-	-	-	-
Profissão (Estudante)	0,146	0,07	0,000649	0,008	0,147	0,286
Profissão (Agricultor)	-0,428	0,07	0,000596	-0,565	-0,427	-0,29
Profissão (Trab. doméstico)	0,382	0,074	0,000627	0,234	0,383	0,529
Profissão (Outras)	-0,095	0,073	0,00065	-0,239	-0,095	0,048
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-0,697	0,039	0,000341	-0,775	-0,697	-0,621
Motivo (Lavar roupa)	-1,45	0,056	0,00046	-1,561	-1,449	-1,341
Motivo (Pescar)	-1,974	0,066	0,000564	-2,104	-1,973	-1,844
Motivo (Nadar)	-0,44	0,047	0,000428	-0,534	-0,44	-0,346
Motivo (Todos os anteriores)	-1,633	0,076	0,000673	-1,786	-1,632	-1,485
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	0,089	0,031	0,0003	0,026	0,089	0,151
Província (Kwanza Sul)	0,351	0,049	0,000456	0,252	0,351	0,447
Naturalidade (Luanda Bengo)	a)	-	-	-	-	-
Naturalidade (Bié Huambo Moxico)	0,354	0,034	0,000305	0,285	0,354	0,422
Naturalidade (Norte)	-0,496	0,047	0,00042	-0,59	-0,495	-0,404
Naturalidade (Sul)	0,54	0,041	0,000361	0,458	0,54	0,62

Tabela 8.2: Estatísticas Modelo Poisson GLM

8.2.2 Estatísticas Binomial Negativa GLM

Binomial Negativa GLM						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	3,183	0,534	0,0054	2,168	3,179	4,243
Género(F)	a)	-	-	-	-	-
Género (M)	0,3988	0,307	0,0034	-0,202	0,391	0,999
Idade	-0,01713	0,01	0,0001	-0,036	-0,017	0,004
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	—	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	—	-	-	-	-	-
Contacto com água (Rio)	a)	-	-	-	-	-
Contacto com água (Lagoa)	-0,5597	0,393	0,0034	-1,29	-0,569	0,239
Contacto com água (Tanque)	-1,292	0,31	0,0036	-1,905	-1,288	-0,685
Conhecimento da doença (Sim)	a)	-	-	-	-	-
Conhecimento da doença (Não)	0,7588	0,348	0,0038	0,082	0,768	1,422
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	1,349	0,311	0,0035	0,757	1,347	1,977
Profissão (Func. Público)	a)	-	-	-	-	-
Profissão (Estudante)	—	-	-	-	-	-
Profissão (Agricultor)	—	-	-	-	-	-
Profissão (Trab. doméstico)	—	-	-	-	-	-
Profissão (Outras)	—	-	-	-	-	-
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-0,8909	0,414	0,0048	-1,687	-0,903	-0,06
Motivo (Lavar roupa)	-0,7472	0,5	0,0057	-1,674	-0,761	0,277
Motivo (Pescar)	-1,421	0,452	0,005	-2,297	-1,432	-0,521
Motivo (Nadar)	-0,6223	0,611	0,0073	-1,757	-0,637	0,674
Motivo (Todos os anteriores)	-1,667	0,6	0,0065	-2,784	-1,684	-0,408
r	0,2823	0,024	0,0002	0,237	0,281	0,331

Tabela 8.3: Estatísticas Modelo Binomial Negativa GLM

8.2.3 Estatísticas Modelo ZIP

ZIP Zeros						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	-0,9151	0,435	0,0039	-1,765	-0,9123	-0,0611
Género(F)	a)	-	-	-	-	-
Género (M)	0,1554	0,2882	0,0025	-0,4077	0,1554	0,7191
Idade	0,0007145	0,0112	0	-0,0219	0,0008	0,0222
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	.	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	-	-	-	-	-	-
Contacto com água (Rio)	a)	-	-	-	-	-
Contacto com água (Lagoa)	1,837	0,5004	0,0043	0,8837	1,831	2,852
Contacto com água (Tanque)	0,3741	0,3865	0,0035	-0,3937	0,3764	1,131
Conhecimento da doença (Sim)	a)	-	-	-	-	-
Conhecimento da doença (Não)	-	-	-	-	-	-
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	-1,233	0,4529	0,0044	-2,166	-1,216	-0,3935
Profissão (Func. Público)	a)	-	-	-	-	-
Profissão (Estudante)	-	-	-	-	-	-
Profissão (Agricultor)	-	-	-	-	-	-
Profissão (Trab. doméstico)	-	-	-	-	-	-
Profissão (Outras)	-	-	-	-	-	-
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-	-	-	-	-	-
Motivo (Lavar roupa)	-	-	-	-	-	-
Motivo (Pescar)	-	-	-	-	-	-
Motivo (Nadar)	-	-	-	-	-	-
Motivo (Todos os anteriores)	-	-	-	-	-	-
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	-0,3143	0,4165	0,0036	-1,143	-0,3068	0,4955
Província (Kwanza Sul)	-1,46	0,5319	0,0046	-2,545	-1,44	-0,4727
Naturalidade (Luanda Bengo)	a)	-	-	-	-	-
Naturalidade (Bié Huambo Moxico)	-	-	-	-	-	-
Naturalidade (Norte)	-	-	-	-	-	-
Naturalidade (Sul)	-	-	-	-	-	-

ZIP Média						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	1,772	0,1	0,0009	1,574	1,772	1,966
Género(F)	a)	-	-	-	-	-
Género (M)	0,8784	0,042	0,0004	0,796	0,877	0,96
Idade	-0,0144	0,001	0	-0,017	-0,014	-0,011
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	-	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	0,9469	0,046	0,0004	0,856	0,946	1,039
Contacto com água (Rio)	a)	-	-	-	-	-
Contacto com água (Lagoa)	0,1633	0,065	0,0006	0,034	0,163	0,289
Contacto com água (Tanque)	-1,628	0,049	0,0004	-1,724	-1,628	-1,533
Conhecimento da doença (Sim)	a)	-	-	-	-	-
Conhecimento da doença (Não)	1,269	0,045	0,0004	1,182	1,269	1,358
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	1,116	0,031	0,0002	1,054	1,117	1,178
Profissão (Func. Público)	a)	-	-	-	-	-
Profissão (Estudante)	0,3624	0,072	0,0007	0,222	0,362	0,507
Profissão (Agricultor)	-0,2351	0,071	0,0008	-0,375	-0,235	-0,093
Profissão (Trab.doméstico)	0,6893	0,074	0,0008	0,544	0,688	0,836
Profissão (Outras)	0,2561	0,074	0,0008	0,11	0,255	0,403
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-0,9026	0,041	0,0003	-0,983	-0,902	-0,821
Motivo (Lavar roupa)	-1,755	0,056	0,0005	-1,865	-1,755	-1,646
Motivo (Pescar)	-2,246	0,069	0,0006	-2,385	-2,246	-2,112
Motivo (Nadar)	-0,2587	0,053	0,0005	-0,363	-0,258	-0,154
Motivo (Todos os anteriores)	-1,956	0,082	0,0007	-2,117	-1,956	-1,797
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	-0,1127	-	-	-	-	-
Província (Kwanza Sul)	0,1661	0,055	0,0005	0,059	0,165	0,274
Naturalidade (Luanda Bengo)	a)	-	-	-	-	-
Naturalidade (Bié Huambo Moxico)	0,6097	0,037	0,0003	0,538	0,609	0,683
Naturalidade (Norte)	-0,6078	0,049	0,0004	-0,704	-0,608	-0,509
Naturalidade (Sul)	0,3138	0,043	0,0004	0,229	0,313	0,398

Tabela 8.4: Estatísticas Modelo ZIP

8.2.4 Estatísticas Modelo ZIBN

ZIBN Zeros						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	-238,1	354,5	1,724	-1022	-200,2	376,6
Género(F)	a)	-	-	-	-	-
Género (M)	-29,54	236	1,132	-501,9	-28,4	442,6
Idade	-6,471	11,39	0,054	-33,78	-4,493	11,45
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	—	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	—	-	-	-	-	-
Cont. com água (Rio)	a)	-	-	-	-	-
Cont. com água (Lagoa)	—	-	-	-	-	-
Cont. com água (Tanque)	—	-	-	-	-	-
Conhec. da doença (Sim)	a)	-	-	-	-	-
Conhec. da doença (Não)	—	-	-	-	-	-
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	—	-	-	-	-	-
Prof. (Func. Público)	a)	-	-	-	-	-
Prof. (Estudante)	—	-	-	-	-	-
Prof. (Agricultor)	—	-	-	-	-	-
Prof. (Trab. doméstico)	—	-	-	-	-	-
Prof. (Outras)	—	-	-	-	-	-
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-	-	-	-	-	-
Motivo (Lavar roupa)	-	-	-	-	-	-
Motivo (Pescar)	-	-	-	-	-	-
Motivo (Nadar)	-	-	-	-	-	-
Motivo (Todos os anteriores)	-	-	-	-	-	-
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	-	-	-	-	-	-
Província (Kwanza Sul)	-	-	-	-	-	-
Naturalidade (Luanda Bengo)	a)	—	—	—	—	—
Naturalidade (Bié Huambo Moxico)	-	-	-	-	-	-
Naturalidade (Norte)	-	-	-	-	-	-
Naturalidade (Sul)	-	-	-	-	-	-
ZIBN Média						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	2,652	0,596	0,002	1,49	2,648	3,828
Género(F)	a)	-	-	-	-	-
Género (M)	0,4067	0,301	0,001	-0,179	0,405	1,001
Idade	-0,01774	0,01	0	-0,037	-0,017	0,002
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	—	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	0,5272	0,28	0,001	-0,04	0,531	1,061
Cont. com água (Rio)	a)	-	-	-	-	-
Cont. com água (Lagoa)	-0,4677	0,389	0,001	-1,203	-0,477	0,319
Cont. com água (Tanque)	-1,263	0,314	0,001	-1,871	-1,264	-0,636
Conhec. da doença (Sim)	a)	-	-	-	-	-
Conhec. da doença (Não)	0,8873	0,343	0,001	0,192	0,892	1,545
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	1,296	0,312	0,001	0,701	1,29	1,921
Prof. (Func. Público)	a)	-	-	-	-	-
Prof. (Estudante)	—	-	-	-	-	-
Prof. (Agricultor)	—	-	-	-	-	-
Prof. (Trab. doméstico)	—	-	-	-	-	-
Prof. (Outras)	—	-	-	-	-	-
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-0,8882	0,404	0,002	-1,658	-0,897	-0,066
Motivo (Lavar roupa)	-0,6534	0,493	0,002	-1,581	-0,666	0,364
Motivo (Pescar)	-1,399	0,444	0,002	-2,255	-1,407	-0,509
Motivo (Nadar)	-0,6821	0,613	0,003	-1,829	-0,704	0,585
Motivo (Todos os anteriores)	-1,62	0,579	0,002	-2,703	-1,639	-0,428
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	-	-	-	-	-	-
Província (Kwanza Sul)	-	-	-	-	-	-
Naturalidade (Luanda Bengo)	a)	—	—	—	—	—
Naturalidade (Bié Huambo Moxico)	-	-	-	-	-	-
Naturalidade (Norte)	-	-	-	-	-	-
Naturalidade (Sul)	-	-	-	-	-	-
r	0,285	0,024	0	0,24	0,284	0,335

Tabela 8.5: Estatísticas Modelo ZIBN

8.2.5 Estatísticas Modelo ZAP

ZAP Zeros						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	-0,9776	0,418	0,004	-1,792	-0,974	-0,17
Género(F)	a)	-	-	-	-	-
Género (M)	0,2074	0,28	0,002	-0,344	0,206	0,76
Idade	0,00465	0,01	0	-0,015	0,004	0,025
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	—	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	—	-	-	-	-	-
Contacto com água (Rio)	a)	-	-	-	-	-
Contacto com água (Lagoa)	1,881	0,475	0,005	0,966	1,874	2,841
Contacto com água (Tanque)	0,3698	0,367	0,004	-0,367	0,37	1,084
Conhecimento da doença (Sim)	a)	-	-	-	-	-
Conhecimento da doença (Não)	—	-	-	-	-	-
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	-1,295	0,443	0,005	-2,187	-1,273	-0,456
Profissão (Func. Público)	a)	-	-	-	-	-
Profissão (Estudante)	-	-	-	-	-	-
Profissão (Agricultor)	-	-	-	-	-	-
Profissão (Trab. doméstico)	-	-	-	-	-	-
Profissão (Outras)	-	-	-	-	-	-
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-	-	-	-	-	-
Motivo (Lavar roupa)	-	-	-	-	-	-
Motivo (Pescar)	-	-	-	-	-	-
Motivo (Nadar)	-	-	-	-	-	-
Motivo (Todos os anteriores)	-	-	-	-	-	-
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	-0,4125	0,403	0,004	-1,208	-0,413	0,381
Província (Kwanza Sul)	-1,407	0,497	0,005	-2,412	-1,397	-0,463
Naturalidade (Luanda Bengo)	a)	-	-	-	-	-
Naturalidade (Bié Huambo Moxico)	-	-	-	-	-	-
Naturalidade (Norte)	-	-	-	-	-	-
Naturalidade (Sul)	-	-	-	-	-	-
ZAP Média						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	1,765	0,1	0,001	1,569	1,765	1,963
Género(F)	a)	-	-	-	-	-
Género (M)	0,882	0,041	0,0004	0,8	0,882	0,962
Idade	-0,01442	0,001	0	-0,017	-0,014	-0,011
Água Canal. (Sim)	a)	-	-	-	-	-
Água Canal. (Não)	—	-	-	-	-	-
WC (Dentro)	a)	-	-	-	-	-
WC (Fora)	0,9494	0,046	0,0005	0,858	0,949	1,042
Contacto com água (Rio)	a)	-	-	-	-	-
Contacto com água (Lagoa)	0,1647	0,065	0,0006	0,036	0,165	0,294
Contacto com água (Tanque)	-1,63	0,048	0,0005	-1,728	-1,63	-1,536
Conhecimento da doença (Sim)	a)	-	-	-	-	-
Conhecimento da doença (Não)	1,273	0,045	0,0005	1,186	1,273	1,361
Hematúria (Negativo)	a)	-	-	-	-	-
Hematúria (Positivo)	1,118	0,031	0,0003	1,055	1,118	1,179
Profissão (Func. Público)	a)	-	-	-	-	-
Profissão (Estudante)	0,3631	0,073	0,0009	0,22	0,362	0,504
Profissão (Agricultor)	-0,235	0,071	0,0009	-0,374	-0,234	-0,097
Profissão (Trab. doméstico)	0,6912	0,074	0,0009	0,545	0,689	0,838
Profissão (Outras)	0,2575	0,074	0,0009	0,111	0,256	0,405
Motivo (Buscar água)	a)	-	-	-	-	-
Motivo(Higiene Pessoal)	-0,9048	0,041	0,0004	-0,985	-0,904	-0,822
Motivo (Lavar roupa)	-1,761	0,056	0,0006	-1,872	-1,761	-1,651
Motivo (Pescar)	-2,255	0,069	0,0007	-2,394	-2,254	-2,121
Motivo (Nadar)	-0,2588	0,053	0,0005	-0,366	-0,258	-0,154
Motivo (Todos os anteriores)	-1,963	0,082	0,0009	-2,127	-1,962	-1,803
Província (Luanda)	a)	-	-	-	-	-
Província (Bengo)	-0,1145	0,033	0,0003	-0,18	-0,114	-0,049
Província (Kwanza Sul)	0,167	0,055	0,0006	0,056	0,167	0,275
Naturalidade (Luanda Bengo)	a)	-	-	-	-	-
Naturalidade (Bié Huambo Moxico)	0,6115	0,036	0,0004	0,539	0,611	0,684
Naturalidade (Norte)	-0,6091	0,049	0,0005	-0,704	-0,609	-0,511
Naturalidade (Sul)	0,3119	0,042	0,0004	0,226	0,312	0,396

Tabela 8.6: Estatísticas Modelo ZAP

8.2.6 Estatísticas Modelo ZANB

ZABN Zeros						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	-0,982	0,417	0,003	-1,796	-0,981	-0,176
Género(F)	a)	—	—	—	—	—
Género (M)	0,209	0,281	0,002	-0,349	0,202	0,754
Idade	0,004	0,01	0	-0,016	0,004	0,024
Água Canal. (Sim)	a)	—	—	—	—	—
Água Canal. (Não)	—	—	—	—	—	—
WC (Dentro)	a)	—	—	—	—	—
WC (Fora)	—	—	—	—	—	—
Cont. com água (Rio)	a)	—	—	—	—	—
Cont. com água (Lagoa)	1,872	0,479	0,004	0,943	1,866	2,823
Cont. com água (Tanque)	0,369	0,371	0,003	-0,338	0,373	1,109
Conhec. da doença (Sim)	a)	—	—	—	—	—
Conhec. da doença (Não)	—	—	—	—	—	—
Hematúria (Negativo)	a)	—	—	—	—	—
Hematúria (Positivo)	-1,292	0,451	0,004	-2,213	-1,272	-0,435
Prof. (Func. Público)	a)	—	—	—	—	—
Prof. (Estudante)	—	—	—	—	—	—
Prof. (Agricultor)	—	—	—	—	—	—
Prof. (Trab. doméstico)	—	—	—	—	—	—
Prof. (Outras)	—	—	—	—	—	—
Motivo (Buscar água)	a)	—	—	—	—	—
Motivo(Higiene Pessoal)	—	—	—	—	—	—
Motivo (Lavar roupa)	—	—	—	—	—	—
Motivo (Pescar)	—	—	—	—	—	—
Motivo (Nadar)	—	—	—	—	—	—
Motivo (Todos os anteriores)	—	—	—	—	—	—
Província (Luanda)	a)	—	—	—	—	—
Província (Bengo)	-0,406	0,403	0,003	-1,197	-0,401	0,38
Província (Kwanza Sul)	-1,401	0,5	0,004	-2,427	-1,388	-0,455
Naturalidade (Luanda Bengo)	a)	—	—	—	—	—
Naturalidade (Bié Huambo Moxico)	—	—	—	—	—	—
Naturalidade (Norte)	—	—	—	—	—	—
Naturalidade (Sul)	—	—	—	—	—	—
ZABN Média						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	2,675	0,445	0,003	1,805	2,673	3,556
Género(F)	a)	—	—	—	—	—
Género (M)	-0,198	0,206	0,001	-0,596	-0,197	0,214
Idade	-0,021	0,007	0	-0,035	-0,021	-0,006
Água Canal. (Sim)	a)	—	—	—	—	—
Água Canal. (Não)	—	—	—	—	—	—
WC (Dentro)	a)	—	—	—	—	—
WC (Fora)	0,711	0,243	0,002	0,226	0,712	1,185
Cont. com água (Rio)	a)	—	—	—	—	—
Cont. com água (Lagoa)	—	—	—	—	—	—
Cont. com água (Tanque)	—	—	—	—	—	—
Conhec. da doença (Sim)	a)	—	—	—	—	—
Conhec. da doença (Não)	0,634	0,278	0,002	0,071	0,634	1,156
Hematúria (Negativo)	a)	—	—	—	—	—
Hematúria (Positivo)	1	0,233	0,002	0,55	0,994	1,476
Prof. (Func. Público)	a)	—	—	—	—	—
Prof. (Estudante)	—	—	—	—	—	—
Prof. (Agricultor)	—	—	—	—	—	—
Prof. (Trab. doméstico)	—	—	—	—	—	—
Prof. (Outras)	—	—	—	—	—	—
Motivo (Buscar água)	a)	—	—	—	—	—
Motivo(Higiene Pessoal)	—	—	—	—	—	—
Motivo (Lavar roupa)	—	—	—	—	—	—
Motivo (Pescar)	—	—	—	—	—	—
Motivo (Nadar)	—	—	—	—	—	—
Motivo (Todos os anteriores)	—	—	—	—	—	—
Província (Luanda)	a)	—	—	—	—	—
Província (Bengo)	—	—	—	—	—	—
Província (Kwanza Sul)	—	—	—	—	—	—
Naturalidade (Luanda Bengo)	a)	—	—	—	—	—
Naturalidade (Bié Huambo Moxico)	—	—	—	—	—	—
Naturalidade (Norte)	—	—	—	—	—	—
Naturalidade (Sul)	—	—	—	—	—	—
r	0,508	0,043	0	0,429	0,508	0,597

Tabela 8.7: Estatísticas Modelo ZABN

8.2.7 Estatísticas Modelo Binomial Negativa Sobre Parametrizada

Binomial Negativa Sobre Parametrizada - Dispersão						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	3,94	0,8047	0,009932	2,394	3,926	5,576
Género(F)	a)	—	—	—	—	—
Género (M)	-1,552	0,6865	0,009391	-2,944	-1,532	-0,2677
Idade	-1,566	0,6873	0,008944	-2,947	-1,555	-0,2734
Água Canal. (Sim)	a)	—	—	—	—	—
Água Canal. (Não)	—	—	—	—	—	—
WC (Dentro)	a)	—	—	—	—	—
WC (Fora)	-0,8782	0,6771	0,009574	-2,213	-0,8608	0,4114
Cont. com água (Rio)	a)	—	—	—	—	—
Cont. com água (Lagoa)	-1,342	0,3707	0,004568	-2,059	-1,346	-0,5936
Cont. com água (Tanque)	-0,8713	0,4619	0,006367	-1,749	-0,8778	0,08192
Conhec. da doença (Sim)	a)	—	—	—	—	—
Conhec. da doença (Não)	-1,277	0,4385	0,005474	-2,132	-1,274	-0,4174
Hematúria (Negativo)	a)	—	—	—	—	—
Hematúria (Positivo)	-0,7065	0,8472	0,01162	-2,18	-0,7706	1,188
Prof. (Func. Público)	a)	—	—	—	—	—
Prof. (Estudante)	-0,01124	0,01084	0,0001254	-0,03235	-0,01132	0,01084
Prof. (Agricultor)	-1,331	0,4976	0,007119	-2,295	-1,341	-0,3371
Prof. (Trab. doméstico)	0,06391	0,3403	0,004136	-0,6086	0,0658	0,7436
Prof. (Outras)	0,7238	0,2452	0,003155	0,2299	0,7242	1,197
Motivo (Buscar água)	a)	—	—	—	—	—
Motivo(Higiene Pessoal)	-0,8041	0,4019	0,005028	-1,544	-0,8201	0,02291
Motivo (Lavar roupa)	-1,553	0,3365	0,004724	-2,214	-1,551	-0,9
Motivo (Pescar)	1,098	0,3458	0,004229	0,4002	1,103	1,754
Motivo (Nadar)	1,294	0,318	0,003924	0,6824	1,288	1,916
Motivo (Todos os anteriores)	-1,95	0,6538	0,009313	-3,285	-1,919	-0,738
Província (Luanda)	a)	—	—	—	—	—
Província (Bengo)	—	—	—	—	—	—
Província (Kwanza Sul)	—	—	—	—	—	—
Naturalidade (Luanda Bengo)	a)	—	—	—	—	—
Naturalidade (Bié Huambo Moxico)	—	—	—	—	—	—
Naturalidade (Norte)	—	—	—	—	—	—
Naturalidade (Sul)	—	—	—	—	—	—

Binomial Negativa Sobre Parametrizada - Média						
Covariável	Média	Desv. Padrão	Erro MC	Qt. 2.5 %	Mediana	Qt. 97.5 %
Ordenada na Origem	-0,8423	0,331	0,004633	-1,499	-0,8354	-0,1938
Género(F)	a)	—	—	—	—	—
Género (M)	-0,7729	0,261	0,003884	-1,298	-0,7723	-0,264
Idade	-0,6939	0,3264	0,00356	-1,345	-0,6862	-0,05196
Água Canal. (Sim)	a)	—	—	—	—	—
Água Canal. (Não)	—	—	—	—	—	—
WC (Dentro)	a)	—	—	—	—	—
WC (Fora)	0,1062	0,2441	0,003061	-0,3699	0,1079	0,5863
Cont. com água (Rio)	a)	—	—	—	—	—
Cont. com água (Lagoa)	0,886	0,3259	0,003629	0,2426	0,8901	1,524
Cont. com água (Tanque)	0,6352	0,3291	0,004377	-0,0008858	0,638	1,286
Conhec. da doença (Sim)	a)	—	—	—	—	—
Conhec. da doença (Não)	—	—	—	—	—	—
Hematúria (Negativo)	a)	—	—	—	—	—
Hematúria (Positivo)	—	—	—	—	—	—
Prof. (Func. Público)	a)	—	—	—	—	—
Prof. (Estudante)	—	—	—	—	—	—
Prof. (Agricultor)	—	—	—	—	—	—
Prof. (Trab. doméstico)	—	—	—	—	—	—
Prof. (Outras)	—	—	—	—	—	—
Motivo (Buscar água)	a)	—	—	—	—	—
Motivo(Higiene Pessoal)	0,9773	0,3521	0,003768	0,2968	0,9726	1,668
Motivo (Lavar roupa)	-0,08614	0,4577	0,005185	-1,007	-0,0775	0,7752
Motivo (Pescar)	0,9116	0,4297	0,005124	0,04835	0,9162	1,764
Motivo (Nadar)	-0,3561	0,2291	0,002946	-0,8027	-0,3505	0,08041
Motivo (Todos os anteriores)	0,003657	0,007131	0,00009083	-0,01025	0,003715	0,01747
Província (Luanda)	a)	—	—	—	—	—
Província (Bengo)	—	—	—	—	—	—
Província (Kwanza Sul)	—	—	—	—	—	—
Naturalidade (Luanda Bengo)	a)	—	—	—	—	—
Naturalidade (Bié Huambo Moxico)	—	—	—	—	—	—
Naturalidade (Norte)	—	—	—	—	—	—
Naturalidade (Sul)	—	—	—	—	—	—

Tabela 8.8: Estatísticas Modelo Binomial Negativo Sobre Parametrizado

8.3 Exemplos da Análise de Convergência das Cadeias

São apresentadas aqui representações gráficas da *posteriori* dos parâmetros dos modelos, das funções de autocorrelação, que permitem aferir o nível de autocorrelação da amostra obtida, o gráfico Brooks-Gelman-Rubin, que representa a convergência do *potential scale reduction* (ou $\hat{R}$) para 1 e uma representação do traço da cadeia aleatória gerada. Dado o grande número de parâmetros em cada modelo não foi considerado adequado mostrar neste documento todas as representações gráficas associadas à análise das cadeias geradas. Deste modo, como exemplo, apresentam-se apenas os resultados do modelo ZIBN. No suporte digital que acompanha este relatório estão incluídas todas as análises.

8.3.1 ZI Binomial Negativa

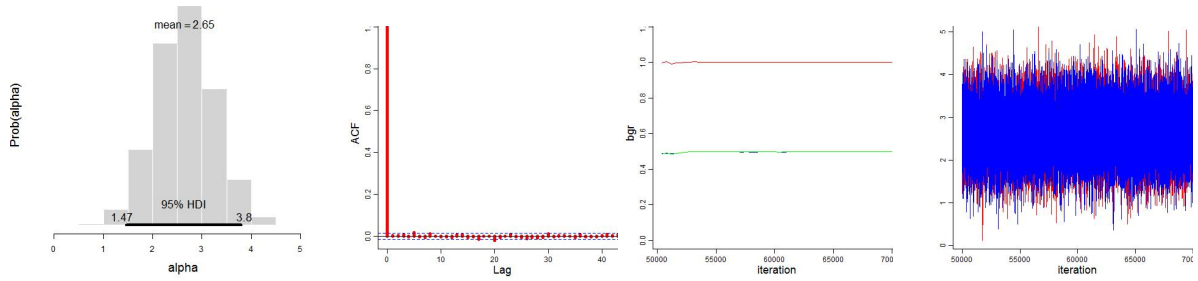


Figura 8.1: Análise de convergência ZI Binomial Negativa - Ordenada na Origem

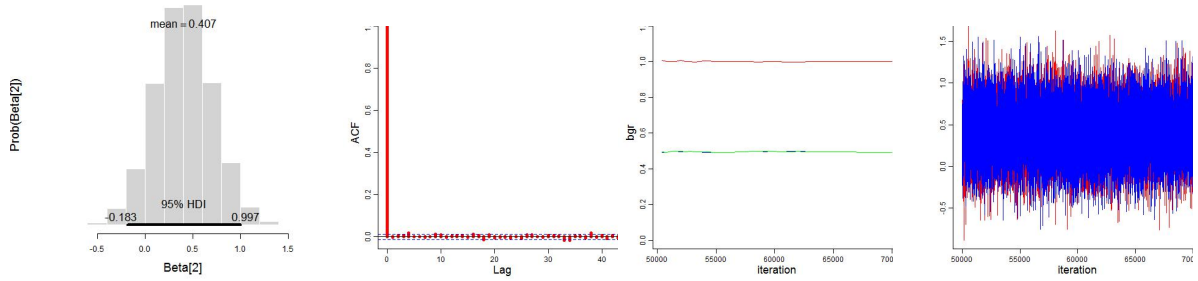


Figura 8.2: Análise de convergência ZI Binomial Negativa - Género: Masculino

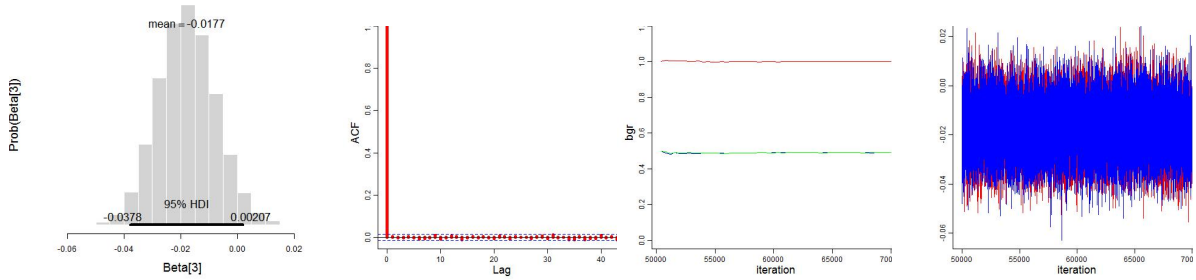


Figura 8.3: Análise de convergência ZI Binomial Negativa - Idade

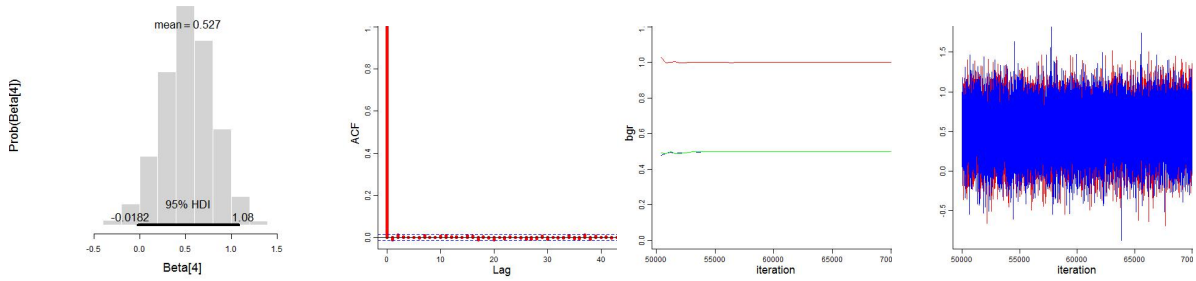


Figura 8.4: Análise de convergência ZI Binomial Negativa - Água Canalizada: Sim

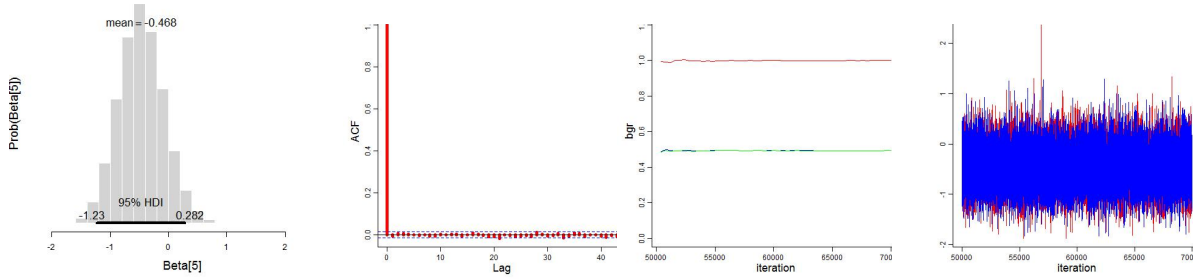


Figura 8.5: Análise de convergência ZI Binomial Negativa - Contacto com água: Lagoa

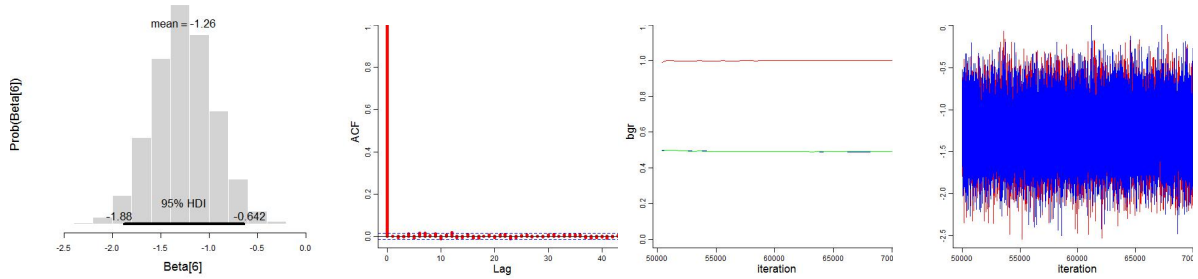


Figura 8.6: Análise de convergência ZI Binomial Negativa - Contacto com água: Tanque

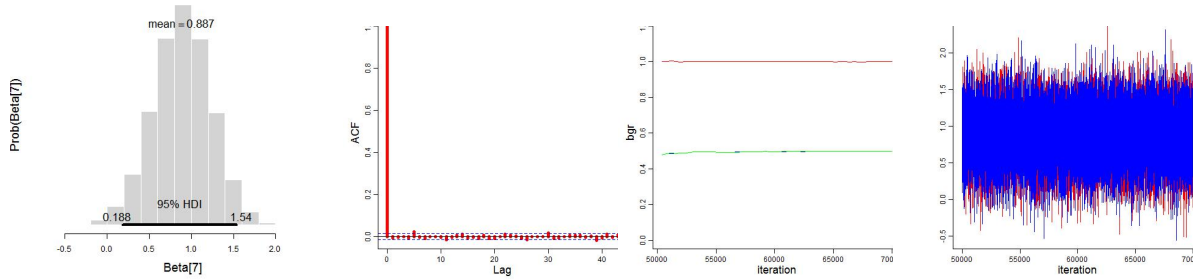


Figura 8.7: Análise de convergência ZI Binomial Negativa - Conhecimento da Doença: Positivo

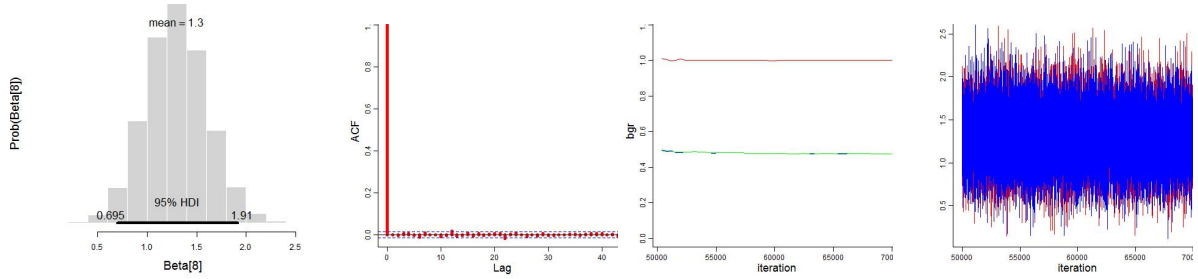


Figura 8.8: Análise de convergência ZI Binomial Negativa - Teste Hematúria: Positivo

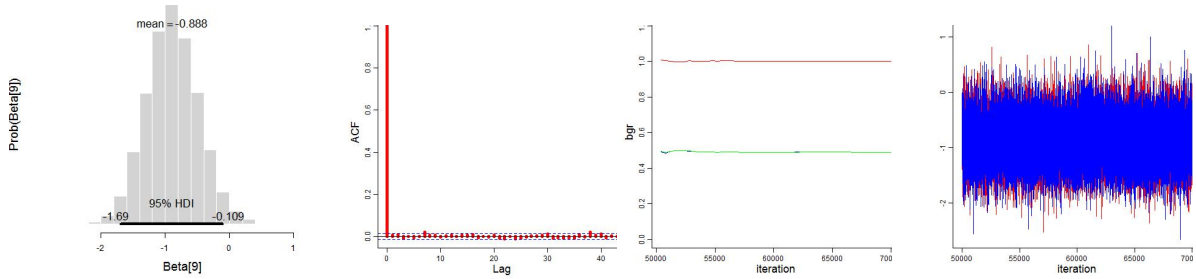


Figura 8.9: Análise de convergência ZI Binomial Negativa - Motivo de contacto com água: Higiene Pessoal

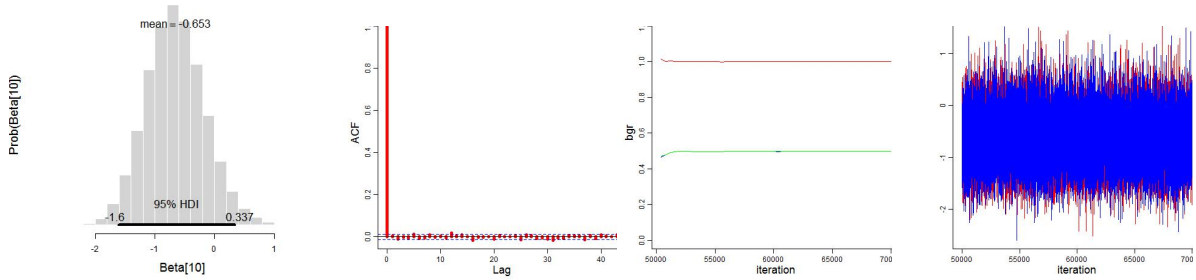


Figura 8.10: Análise de convergência ZI Binomial Negativa - Motivo de contacto com água: Lavar Roupa

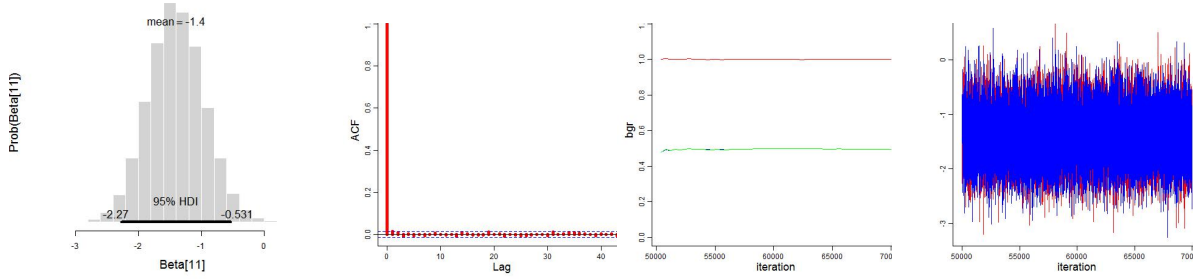


Figura 8.11: Análise de convergência ZI Binomial Negativa - Motivo de contacto com água: Pescar

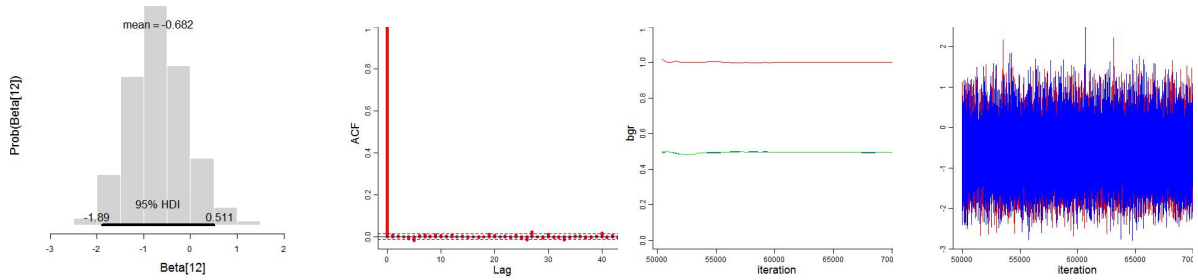


Figura 8.12: Análise de convergência ZI Binomial Negativa - Motivo de contacto com água: Nadar

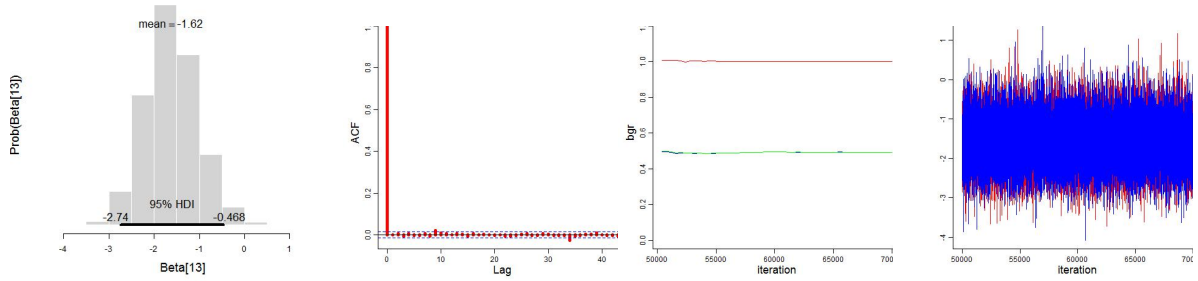


Figura 8.13: Análise de convergência ZI Binomial Negativa - Motivo de contacto com água: Todos os anteriores

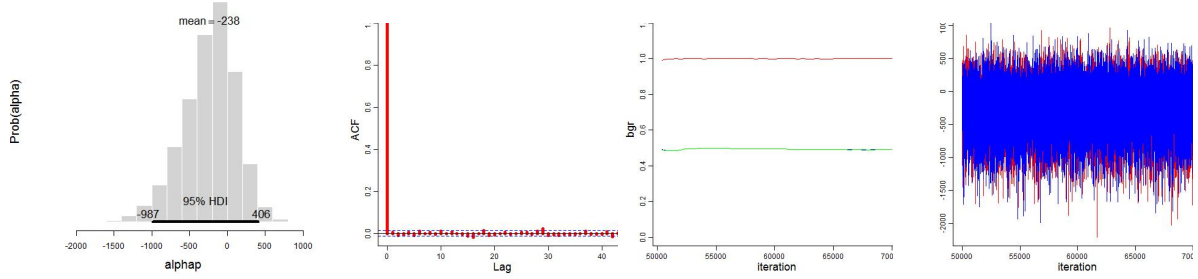


Figura 8.14: Análise de convergência ZI Binomial Negativa - Zeros - Ordenada na Origem

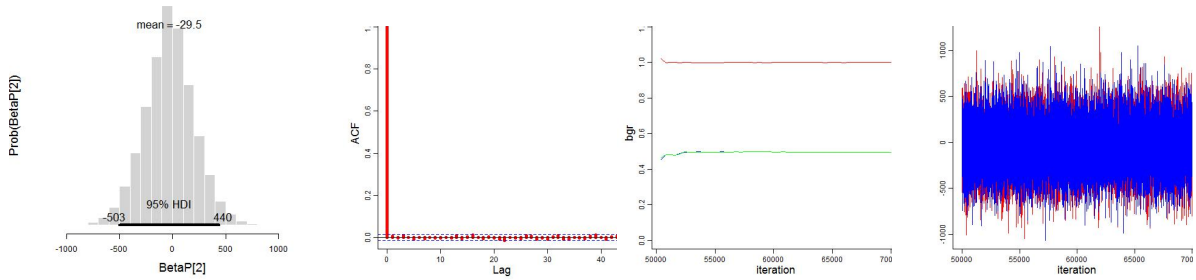


Figura 8.15: Análise de convergência ZI Binomial Negativa - Zeros - Género: Masculino

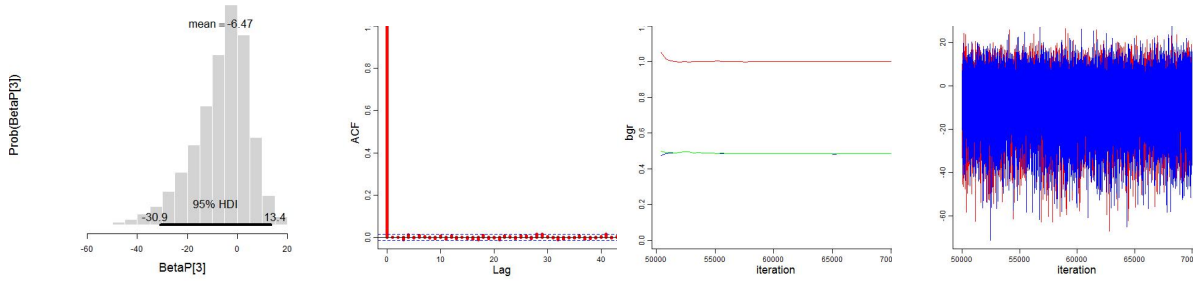


Figura 8.16: Análise de convergência ZI Binomial Negativa - Zeros - Idade

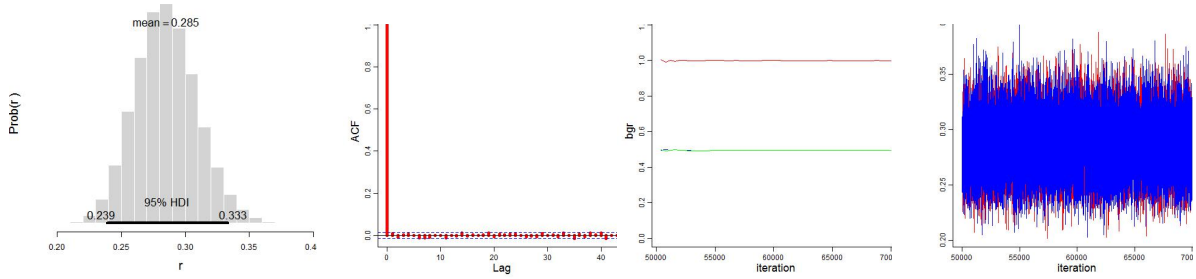


Figura 8.17: Análise de convergência ZIBN - Parâmetro de Dispersão : r

8.4 Definição de uma distribuição arbitrária em Bugs - *Zeros-Ones Trick*

Casos complexos de distribuições como as usadas em vários modelos neste trabalho não existem no OPENBUGS. De forma de conseguir implementá-los usa-se uma técnica chamada Zeros-Ones Trick (Ntzoufras 2009 e Tricks: Advanced Use of the BUGS Language).

É possível recorrer à distribuição Poisson para especificar arbitrariamente a verosimilhança de um modelo em BUGS.
Assuma-se que um modelo tem log-verosimilhança $l_i = log(p(y_i \mid \theta))$, assim podemos escrever a verosimilhança do modelo como:

$$L(\theta \mid y) = \prod_{i=1}^{n} e^{l_i} = \prod_{i=1}^{n} \frac{e^{-(-l_i)}(-l_i)^0}{0!} = \prod_{i=1}^{n} p_{poisson}(0 \mid -l_i).$$

A verosimilhança pode ser escrita como o produto de pseudo variáveis aleatórias com distribuição Poisson, com média $-l_i$, em que os valores observados são zero. Para garantir que as médias das variáveis aleatórias de Poisson são positivas soma-se uma constante K suficientemente grande de modo a que $K - l_i > 0$ (Ntzoufras, 2009).

Em BUGS, para uma amostra de dimensão n, um exemplo de programa usado para implementar o *Zeros-Ones Trick* será:

```
K < - 10000
for (i in 1:n)
{
Zeros[i] < - 0
Zeros[i] ~ dpois(Zeros.mean[i])
Zeros.mean[i] <-  -l[i]+K
l(i)<- (Verosimilhança do Modelo Escolhido)
}
```

Esta técnica também pode usada de forma análoga com a distribuição Bernoulli (detalhes sobre a sua implementação podem ser encontrada em Ntzoufras (2009)).

8.5 Exemplos Programas R - Bugs

O conjunto de programas criado é bastante extenso, pelo que listar todo o código gerado não é seria possível numa versão impressa. Foram colocados aqui apenas os programas usados no modelo ZIBN, os programas auxiliares e a respectiva análise descritiva. No suporte digital que acompanha este relatório está incluído todo o código.

8.5.1 Programa ZIBN

```
graphics.off()
 rm(list=ls(all=TRUE))
 library(BRugs)

 #--------------------------------------------------------------------------------
 # Modelo.

getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2") #Preparar aqui a pasta


modelString =
"model
{
# Variable Names:
# ovos10ml # profissao # idade # sexo # hematuria # saberschist
# motivo # h2o_canalizada # wc #  contacto_h2o  # provincia # local natura[]

K<-10000

# Creating the dummies
for(i in 1:n)
{

#*****************  Género ( 0 = Fem ; 1= Masc )  **********************
sexo_2[i]<-equals(sexo[i],1)              # Mulher é Base

#**************************  Agua Calizada (0= Não tem; 1 = Tem água ) **************************
h2o_canal_2[i]<-equals(h2o_canalizada[i] ,0)      # o Sim é Base

#***************************Local WC ( 1= Fora de Casa ; 2 = Dentro )  ************************************
wc_2[i] <-equals(wc[i] ,1) # Dentro de casa é Base

#*****************************  Contacto com a água ( 1= Rio ; 2 = Lagoa  ; 3 = Tanque)  *************************

contacto_2[i]<-equals(contacto_h2o[i],2)  # Rio é Base
contacto_3[i]<-equals(contacto_h2o[i],3)

#Construção de Var dicotomicas
contacto_novo_2[i]<-equals(contacto_2[i],1)
contacto_novo_3[i]<-equals(contacto_3[i],1)

#*****************************     Mac ( 0= Negativo ; 1 = Positivo  )  ******************************************
hematuria_2[i]<-equals( hematuria[i] ,1)   # Negativo é Base
#*****************************  Conhecimento da Doença ( 0= Não sabe ; 1 = Sabe  ) *********************************
saberschist_2[i]<-equals( saberschist[i]  ,0)   # Sim é Base
#***************** Profissão ( 1= Agricultor ; 2 = F. Publico ; 3 =Trab. Doméstico ; 4= Estudante ; 5=Outros ) *****************
profissao_2[i]<-equals(profissao[i],1) # F. Publico é Base
profissao_3[i]<-equals(profissao[i],3)
profissao_4[i]<-equals(profissao[i],4)
profissao_5[i]<-equals(profissao[i],5)

#Construção de Var dicotomicas
```

```
profissao_novo_2[i]<-equals(profissao_2[i],1)
profissao_novo_3[i]<-equals(profissao_3[i],1)
profissao_novo_4[i]<-equals(profissao_4[i],1)
profissao_novo_5[i]<-equals(profissao_5[i],1)

#Motivo de Contacto com a água
#( 1= Busc. Agua ; 2 = Pescar ; 3 = Lav. Roupa ; 4= Hig. Pessoal ; 5=Nadar ; 6=Outros)
motivo_2[i]<-equals(motivo[i],2)  #Busc. Agua é Base
motivo_3[i]<-equals(motivo[i],3)
motivo_4[i]<-equals(motivo[i],4)
motivo_5[i]<-equals(motivo[i],5)
motivo_6[i]<-equals(motivo[i],6)

# Construção de Var dicotomicas
motivo_novo_2[i]<-equals(motivo_2[i],1)
motivo_novo_3[i]<-equals(motivo_3[i],1)
motivo_novo_4[i]<-equals(motivo_4[i],1)
motivo_novo_5[i]<-equals(motivo_5[i],1)
motivo_novo_6[i]<-equals(motivo_6[i],1)

#**********************   Provincia ( 1 = Luanda ; 2= Bengo ; 3 =K. Sul ) ******************

provincia_2[i]<-equals(provincia[i],2) #Luanda é Base
provincia_3[i]<-equals(provincia[i],3)

# Construção de Var dicotomicas

provincia_novo_2[i]<-equals(provincia_2[i],1)
provincia_novo_3[i]<-equals(provincia_3[i],1)

#********** Naturalidade ( 1 = Luanda; Bengo ; 2= Bié, Huambo,Moxico ; 3 =Norte  ; 4 =Sul) ******

natura_2[i]<-equals(natura[i],2)       #Luanda; Bengo é base
natura_3[i]<-equals(natura[i],3)
natura_4[i]<-equals(natura[i],4)

# Construção de Var dicotomicas

natura_novo_2[i]<-equals(natura_2[i],1)
natura_novo_3[i]<-equals(natura_3[i],1)
natura_novo_4[i]<-equals(natura_4[i],1)

d[i]<- equals(ovos10ml[i], 0)

}

for(i in 1:n) # Preparar os Valores Médio para análise
{

sexo_mean[i]    <- ( sexo_2[i] - mean(sexo_2[] ) )
idade_mean[i]   <-( idade[i] - mean(idade[] )  )
h2o_canal_mean[i]   <- ( h2o_canal_2[i] - mean(h2o_canal_2[])  )
wc_mean[i]   <- ( wc_2[i] - mean(wc_2[])  )
contacto_mean_2[i]   <- (contacto_novo_2[i]   - mean(contacto_novo_2[])  )
contacto_mean_3[i]   <-(contacto_novo_3[i] - mean(contacto_novo_3[])   )
saberschist_mean[i]   <- ( saberschist_2[i] - mean(saberschist_2[])  )
hematuria_mean[i]   <-( hematuria_2[i] - mean(hematuria_2[])  )
profissao_mean_2[i]   <- ( profissao_novo_2[i]  -  mean(profissao_novo_2[])  )
profissao_mean_3[i]   <- ( profissao_novo_3[i] -  mean( profissao_novo_3[])  )
profissao_mean_4[i]   <-  ( profissao_novo_4[i] - mean(profissao_novo_4[])  )
profissao_mean_5[i]   <-  ( profissao_novo_5[i] - mean(profissao_novo_5[])  )
motivo_mean_2[i]   <- ( motivo_novo_2[i]  - mean( motivo_novo_2[]) )
motivo_mean_3[i]   <- ( motivo_novo_3[i] - mean(motivo_novo_3[])   )
motivo_mean_4[i]  <- ( motivo_novo_4[i]  - mean(motivo_novo_4[])   )
motivo_mean_5[i]   <- ( motivo_novo_5[i]  - mean( motivo_novo_5[])  )
motivo_mean_6[i]   <-  ( motivo_novo_6[i]  - mean(motivo_novo_6[])  )
provincia_mean_2[i]   <-  ( provincia_novo_2[i]- mean(provincia_novo_2[] ) )
provincia_mean_3[i]   <-  ( provincia_novo_3[i] - mean(provincia_novo_3[] ) )
natura_mean_2[i]   <- ( natura_novo_2[i]- mean(natura_novo_2[]  ))
```

```
natura_mean_3[i]   <- ( natura_novo_3[i] - mean(natura_novo_3[] ))
natura_mean_4[i]   <- ( natura_novo_4[i] - mean(natura_novo_4[] ))
}


for(i in 1:n)
{
Zi[i] <- 0
    Zi[i] ~ dpois(phi[i])
    phi[i]<-  - ll[i]+K

#ovos10ml[i] ~ dnegbin( p.ind[i], r )
#p.ind[i] <- r /( r +lambda.ind[i] )


ll[i] <-       d[i] *     log(p0[i] +(1-p0[i])*pow(pstar[i],r))+
(1-d[i])* ( log(1-p0[i])+loggam(ovos10ml[i]+r)- logfact(ovos10ml[i]) - loggam(r) + r*log(pstar[i])
+ ovos10ml[i]* log(1-pstar[i]))

pstar[i] <- r/(r+mu[i])

log(mu[i])     <-     beta[1] +
          beta[2] *  sexo_mean[i]  +
 beta[3] * idade_mean[i]  +
                              beta[4] * wc_mean[i]  +
                              beta[5] *  contacto_mean_2[i]    +
          beta[6] *  contacto_mean_3[i]   +
                              beta[7] *  saberschist_mean[i]   +
                              beta[8] *    hematuria_mean[i]    +
                              beta[9] *   motivo_mean_2[i]    +
                              beta[10] *  motivo_mean_3[i]    +
    beta[11] * motivo_mean_4[i] +
    beta[12] *  motivo_mean_5[i] +
    beta[13] *motivo_mean_6[i]

 logit(p0[i])    <- betap[1] +
          betap[2] *  sexo_mean[i]  +
betap[3] * idade_mean[i]

prob[i]<-exp(ll[i])
icpo[i]<-1/(prob[i]+0.0001)
}

alpha <-         beta[1]- beta[2]* mean(sexo_2[])
- beta[3]* mean(idade[])
- beta[4]* mean(wc_2[])
- beta[5]* mean(contacto_novo_2[])
      - beta[6]* mean(contacto_novo_3[])
- beta[7]* mean(saberschist_2[])
                      - beta[8]* mean(hematuria_2[])
      - beta[9]* mean(motivo_novo_2[])
- beta[10]* mean(motivo_novo_3[])
- beta[11]* mean(motivo_novo_4[])
      - beta[12]* mean(motivo_novo_5[])
- beta[13]* mean(motivo_novo_6[])

alphap <-         betap[1]- betap[2]* mean(sexo_2[])
- betap[3]* mean(idade[])

#  Priors:
for (j in 1:M) {
beta[j] ~ dnorm(0, 0.00001)
}

#  Priors:
for (j in 1:N) {
betap[j] ~ dnorm(0, 0.00001)
}

r ~ dgamma( 0.001, 0.001 )  # Netzoufras
```

8. ANEXOS

```
# r  <- exp(logr)  Jare paper
#logr ~ dnorm(0.0, 0.01)I(-5, 5)

# Beta estimado
for (j in 1:M)
{
   b[j]<-exp(beta[j])
}

for (j in 1:N)
{
   bp[j]<-exp(betap[j])
}

#Mean.Deviance
Deviance <-  (-2*sum(ll[1:n]))
}"

 # Write the modelString to a file, using R commands:
 .temp = file("modelozanegbinomial.txt","w") ; writeLines(modelString,con=.temp) ; close(.temp)
 #-------------------------------------------------------------------------

 modelCheck("modelozanegbinomial.txt") # Check Model

# file.show("modelozanegbinomial.txt")  #Mostrar o Modelo
 #-------------------------------------------------------------------------
 # THE DATA.

  modelData("Dados_bugs.txt")            # Ler os Dados - Foi retirado do Bugs
# file.show("Data_Bugs.txt")   #Mostrar os dados
 #-------------------------------------------------------------------------
 # Inicializar as Cadeias

 nChains = 2
 modelCompile( numChains = nChains ) # Compilar e definir o nr° de cadeias

 modelinits=
"list(r=2, beta=c(0,0,0,0,0,0,0,0,0,0,0,0),betap=c(-700,0,0))"

 # Write the modelString to a file, using R commands:
 .temp = file("inits1.txt","w") ; writeLines(modelinits,con=.temp) ; close(.temp) # Guardar cadeia inicial

modelInits("inits1.txt", chainNum = 1)

 modelinits=
"list(r=1,beta=c(-0.1,2,0,0.1,0,2,0,6,0,0.56,0,0.12,1),betap=c(-1,5,-1))"

 # Write the modelString to a file, using R commands:
 .temp = file("inits2.txt","w") ; writeLines(modelinits,con=.temp) ; close(.temp) # Guardar cadeia inicial

modelInits("inits2.txt", chainNum = 2)

 #modelinits=
#"list(beta=c(2,0.1,0.1,0.1,0.1,0.1,0,0,0,0,0,0.5,0.5,0,0,0.5,0,0,0,0.5,0,0,0),
betap=c(2,0.1,0.1,0.1,0.1,0.1,0,0,0,0,0,0.5,0.5,0,0,0.5,0,0,0,0.5,0,0,0))"

 # Write the modelString to a file, using R commands:
 #.temp = file("inits3.txt","w") ; writeLines(modelinits,con=.temp) ; close(.temp) # Guardar cadeia inicial

#modelInits("inits3.txt", chainNum = 3)

# modelinits=
#"list(beta=c(2,1.1,0.1,5.1,0.1,0.1,0,0,0,0,0,0.5,0.5,0,4,0.5,0,0,3,0.5,0,0,0),
betap=c(2,0.1,0.1,0.1,0.1,3.1,0,0.5,0,4,1,0.5,0.5,0,0,0.5,0,0,0,0.5,0,0,0))"

 # Write the modelString to a file, using R commands:
 #.temp = file("inits4.txt","w") ; writeLines(modelinits,con=.temp) ; close(.temp) # Guardar cadeia inicial
#modelInits("inits4.txt", chainNum = 4)
#-------------------------------------------------------------------------------------------
```

```
 # Correr o Modelo.

 # Definir o Burn in:
 burninSteps = 50000

ini.Burn <- Sys.time()
   modelUpdate( burninSteps )
end.Burn <- Sys.time()

# Quais as Variáveis a seguir e o número de updates :
# samplesSet( c("alpha","beta","alphap","betap","Deviance","r") )
samplesSet( c("icpo") )

nPerChain = 20000
thinning<- 500

ini.update <- Sys.time()
  modelUpdate( nPerChain , thin=thinning)#dá para definir aqui o step
end.update <- Sys.time()


#---------------------------- Guardar a situação da Cadeia  -------------------------------

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Cadeias - Backup") #Preparar aqui a pasta
getwd()

#Guarda os valores para cada parametro obtidos actualmente
stemdir<- paste(getwd(),"/Set", sep = "")
modelSaveState(stem=stemdir)

#Guarda ficheiros Coda das cadeias

stemdir<- paste(getwd(),"/Cadeia", sep = "")
samplesCoda("*", stem=stemdir, beg = samplesGetBeg(),
end = samplesGetEnd(), firstChain = samplesGetFirstChain(),
lastChain = samplesGetLastChain(), thin = samplesGetThin())


setwd(old) #repor a directoria
getwd()

#--------------------Análise de Convergência das Cadeias  --------------

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Cadeias - Backup") #Preparar aqui a pasta
getwd()
cadeia <- read.openbugs(stem=stemdir) # Ler os Ficheiros Coda
setwd(old) #repor a directoria
getwd()

#Produzir as diferentes Diagnósticos de Convergência

#cat('\nGráfico de Evolução das Cadeias\n')
plot(cadeia, trace =TRUE  , density = FALSE, smooth = FALSE,
auto.layout = TRUE, ask = dev.interactive())

cat('\nEffective sample size for estimating the mean\n')
print(effectiveSize(cadeia ))

cat('\nGeweke - diagnostico de Convergencia\n')
print(geweke.diag(cadeia ),pvalue = 0.05)
#geweke.plot(cadeia, auto.layout = TRUE, ask=TRUE) Não está a funcionar bem

cat('\nAutocorrelações\n')
print(round(autocorr.diag(cadeia ),5))
autocorr.plot(cadeia,ask=TRUE,lag.max=10 )

cat('\nHeidelberger and Welch - diagnostico de convergencia\n')
print(heidel.diag(cadeia ))
```

```
cat('\nRaftery and Lewis diagnostic\n')
print(raftery.diag(cadeia ))

if (nchain(cadeia)>1){
   cat('\nGelman and Rubin - diagnostico de convergencia\n')
   print(gelman.diag(cadeia))
   gelman.plot(cadeia,ask=TRUE)

}

cat('\nDensidades\n')
plot(cadeia, trace = FALSE, density = TRUE, smooth = FALSE,
auto.layout = TRUE, ask = dev.interactive())

cat('\nTaxa de Rejeição\n')
rejectionRate(cadeia)
 # ----------------------------Análise com coda ou boa ----------------------------

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Cadeias - Backup") #Preparar aqui a pasta
getwd()

library(coda)
library(boa)

#codamenu()
#boa.menu()

setwd(old) #repor a directoria
getwd()
#--------------------------------------------------------------------------------

# Resultados

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Resultados") #Preparar aqui a pasta
getwd()

Resultados <-samplesStats("*") # the summarized results
#intervalos de Máxima credibilidade coda
hpd <- HPDinterval(cadeia, prob = 0.95) #isto não fica lá muito bem

final <- cbind(Resultados,hpd ) #juntei o intervalo hpd a este output
write.table(final , "ResultadoModelo.txt", sep=" ") # Guardar os Resultados

Resultadosicpo <-samplesStats("*") # the summarized results
write.table(Resultadosicpo , "ResultadoModeloICPO.txt", sep=" ") # Guardar os Resultados

#Mean.Deviance <- Resultados[1,1]
#Mean.Deviance
#Paramaters <- Resultados[2:48,]
#nparameter<- nrow(Paramaters)

#Guardar os parametros
#Betas<-c(Paramaters[1:23,1])
#Betasp<-Paramaters[24:46,1]

#r <-Paramaters[47,1]

## some plots
#samplesHistory("*", mfrow = c(4, 2)) # plot the chain,
#samplesDensity("beta")              # plot the densities,
#samplesBgr("beta[1:1]")            # plot the bgr statistics, and
#samplesAutoC("beta[1]", 1)          # plot autocorrelations of 1st chain

setwd(old) #repor a directoria
getwd()

#--------------------------------------------------------------------------------
```

```
#---------------------------    Calculo do DIC  ----------------------------------------
#-------------------------------------------------------------------------------------

#Importar dados
dados_dic <- read.table("Dados_Dic.txt", header=TRUE) #tabelas com os dados sem o End ao fim do bugs

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Resultados") #Preparar aqui a pasta
getwd()

res <- read.table("ResultadoModelo.txt", header=TRUE) #tabelas com os dados sem o End ao fim do bugs

Mean.Deviance <- res [1,1]
ParamatersP1 <- res [2,1]
ParamatersP2 <- res [3,1]

Paramaters1 <- res [5:16,1]
Paramaters2 <- res [18:19,1]

r <- res [20,1]

Betas <- c(ParamatersP1 ,Paramaters1 )
Betasp <- c(ParamatersP2 ,Paramaters2 )

# Calculo do valor dos parametros para cada individuo

#Inicializar parametros
b_intercept<- NULL
b_sexo<- NULL
b_idade<- NULL
b_h20_canal<- NULL
b_wc<- NULL
b_contacto_h2o<- NULL
b_saberschist<- NULL
b_hematuria<- NULL
b_profissao<- NULL
b_motivo<- NULL
b_provincia<- NULL
b_natura<- NULL

b<- NULL

#Inicializar parametros P
bp_intercept<- NULL
bp_sexo<- NULL
bp_idade<- NULL
bp_h20_canal<- NULL
bp_wc<- NULL
bp_contacto_h2o<- NULL
bp_saberschist<- NULL
bp_hematuria<- NULL
bp_profissao<- NULL
bp_motivo<- NULL
bp_provincia<- NULL
bp_natura<- NULL

bp<- NULL

for(i in 1:nrow(dados_dic))
{
#Intercept
b_intercept=rbind(b_intercept,Betas[1])
bp_intercept=rbind(bp_intercept,Betasp[1])


#Sexo
if ( dados_dic$sexo[i]== 1)
 {b=Betas[2]
  bp=Betasp[2] }
else   {b=0
```

```
  bp=0}
b_sexo=rbind(b_sexo,b)
bp_sexo=rbind(bp_sexo,bp)

#Idade
 b=Betas[3]*dados_dic$idade[i]
 b_idade=rbind(b_idade,b)

 bp=Betasp[3]*dados_dic$idade[i]
 bp_idade=rbind(bp_idade,bp)


#Água Canalizada
#if ( dados_dic$h2o_canalizada[i]== 0)
# {b=Betas[4]
#  bp=Betasp[4]
# }
#else   {b=0
#   bp=0
# }
#b_h20_canal=rbind(b_h20_canal,b)
#bp_h20_canal=rbind(bp_h20_canal,bp)

#Wc
if ( dados_dic$wc[i]== 1)
 {b=Betas[4]
 }
else   {b=0
 }
b_wc=rbind(b_wc,b)


#Contacto com a àgua
if ( dados_dic$contacto_h2o[i]== 2)
 {b=Betas[5]
   }
else if ( dados_dic$contacto_h2o[i]== 3)
    {b=Betas[6]
 }
else   {b=0
 }
b_contacto_h2o=rbind(b_contacto_h2o,b)

#Sabe da Doença
if ( dados_dic$saberschist[i]== 0)
 {b=Betas[7]
  }
else   {b=0
  }
b_saberschist=rbind(b_saberschist,b)

#Mac
if ( dados_dic$hematuria[i]== 1)
 {b=Betas[8]
 }
else   {b=0
 }
b_hematuria=rbind(b_hematuria,b)

#Profissão
#if ( dados_dic$profissao[i]== 1)
# {b=Betas[10]
#  bp=Betasp[10]}
#else if ( dados_dic$profissao[i]== 3)
   # {b=Betas[11]
#  bp=Betasp[11]}
#else if ( dados_dic$profissao[i]== 4)
# {b=Betas[12]
#  bp=Betasp[12]}
#else if ( dados_dic$profissao[i]== 5)
```

```
# {b=Betas[13]
#  bp=Betasp[13]}
#else   {b=0
#   bp=0}
#b_profissao=rbind(b_profissao,b)
#bp_profissao=rbind(bp_profissao,bp)

#Motivo de Contacto com a água
if ( dados_dic$motivo[i]== 2)
 {b=Betas[9]
  }
else if ( dados_dic$motivo[i]== 3)
    {b=Betas[10]
  }
else if ( dados_dic$motivo[i]== 4)
 {b=Betas[11]
  }
else if ( dados_dic$motivo[i]== 5)
 {b=Betas[12]
 }
else if ( dados_dic$motivo[i]== 6)
 {b=Betas[13]
 }
else   {b=0
 }
b_motivo=rbind(b_motivo,b)

#Provincia
#if ( dados_dic$provincia[i]== 2)
# {b=Betas[19]
#  bp=Betasp[19] }
#else if ( dados_dic$provincia[i]== 3)
   # {b=Betas[20]
#  bp=Betasp[20]}
#else   {b=0
#   bp=0}
#b_provincia=rbind(b_provincia,b)
#bp_provincia=rbind(bp_provincia,bp)

#Naturalidade
#if ( dados_dic$natura[i]== 2)
# {b=Betas[21]
#  bp=Betasp[21] }
#else if ( dados_dic$natura[i]== 3)
   # {b=Betas[22]
#  bp=Betasp[22]}
#else if ( dados_dic$natura[i]== 4)
# {b=Betas[23]
#  bp=Betasp[23]}
#else   {b=0
#   bp=0}
#b_natura=rbind(b_natura,b)
#bp_natura=rbind(bp_natura,bp)
}

# Matriz com os parametros para cada individuo

Matrixparameters<- data.frame(b_intercept,b_sexo,b_idade,b_wc, b_saberschist,b_hematuria , b_motivo )
write.table(Matrixparameters, "Resultados_Parametros_individuos.txt", sep=" ")

Matrixparametersp<- data.frame(bp_intercept,bp_sexo,bp_idade)
write.table(Matrixparametersp, "Resultados_Parametros_individuos_p.txt", sep=" ")

#Soma dos parametros para cada individuo e cálculo de Lambda

lambda<- NULL  #Vector de Lambdas
p0<- NULL #Vector de P
pstar<-NULL

for(i in 1:nrow(dados_dic))
```

8. ANEXOS

```
{
lambda<-rbind(lambda,exp(sum(Matrixparameters[i,])))

p0<-rbind(p0,exp(sum(Matrixparametersp[i,])) / (1+exp(sum(Matrixparametersp[i,]))))

pstar<-rbind(pstar,r/(r+exp(sum(Matrixparameters[i,]))))
}

cbind(lambda,p0,pstar)

#Calculo do logaritmo da probabilidade de cada individuo

bnegative_zeros<- NULL
bnegative   <- NULL

for(i in 1:nrow(dados_dic))
{

bnegative_zeros<- NULL
if ( dados_dic$ovos10ml[i]== 0)
 { #Se Zero
 bnegative_zeros <- log (p0[i] + (1-p0[i])*pstar[i]^r )
 }
else
 { #Se maior que Zero
 bnegative_zeros <- log(1-p0[i])+ lgamma(dados_dic$ovos10ml[i]+r)-
 lgamma(dados_dic$ovos10ml[i]+1) - lgamma(r) + r*log(pstar[i]) + dados_dic$ovos10ml[i]* log(1-pstar[i])
 }
bnegative<-rbind(bnegative,bnegative_zeros)
}


#cbind(bnegative,dados_dic$ovos10ml,p0,pstar,lambda)

#----------------------------- Medidas Usuais-----------------------------

numParametros<-length(Betas )+length(Betasp )+length(r)
nrobservacoes<-nrow(dados_dic)

deviance <-  -2*sum(bnegative)
deviance2 <-  -2*sum(bnegative)
DIC <- 2*Mean.Deviance - deviance
AIC <- deviance  + 2* (numParametros)
BIC <- deviance  + 2*(numParametros)*log(nrobservacoes)

#-------------------------------------------------------------------------

#OutPut
tempo.burnin<- as.numeric(difftime(end.Burn,ini.Burn,unit="mins"))
tempo.update<- as.numeric(difftime(end.update,ini.update,unit="mins"))
tempo.update2<- as.numeric(difftime(end.update2,ini.update2,unit="mins"))

tempo.update_f<- tempo.update2 + tempo.update

Dados_modelo <- data.frame(deviance,Mean.Deviance,DIC,AIC,BIC,numParametros,burninSteps, nPerChain, thinning,nChains,tempo.burnin,tempo.update_f)

write.table(Dados_modelo, "Resumo_Modelo.txt", sep=" ")

setwd(old)
getwd()
#-------------------------------------------------------------------------
#----------------------------  Gráficos -------------------------------------
#-------------------------------------------------------------------------

# Verificar Cadeias

 source("plotChains.R")
 source("plotPost.R")

#Construir os gráficos das cadeias e da densidade posterior
```

```
for(i in 2:13) # Numero de linhas da vector dos parametros
{

#Construir os gráficos das cadeias
old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()

param <- paste("beta[", i,"]", sep = "")
plotChains( param  , saveplots=T ,filenameroot="GraficoCadeia" )
graphics.off()

#Repor a directoria
setwd(old)
getwd()

# Verificar Densidades Posteriores com Intervalo HDI

betasample = samplesSample(param )

x <- paste("Beta[", i,"]", sep = "")
py <- paste("Prob(Beta[", i,"])", sep = "")

plotPost( betasample , credMass=0.95, xlab=x , ylab=py )
densidadejpeg <- paste("DensidadeB",i ,".jpeg", sep = "")
densidade <- paste("DensidadeB",i ,".eps", sep = "")
old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()
dev.copy2eps(file=densidade ) #Guardar Densidade Posteriores
dev.copy(jpeg,file=densidadejpeg ) #Guardar Densidade Posteriores

graphics.off()
setwd(old)
getwd()
}

#Gráfico do predictor Linear para a prob Zero

for(i in 2:3 ) # Numero de linhas da vector dos parametros
{
#Contruir os gráficos das cadeias

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()

param <- paste("betap[", i,"]", sep = "")
plotChains( param  , saveplots=T ,filenameroot="GraficoCadeia" )
graphics.off()

#Repor a directoria
setwd(old)
getwd()


# Verificar Densidades Posteriores com Intervalo HDI

betasample = samplesSample(param )

x <- paste("BetaP[", i,"]", sep = "")
py <- paste("Prob(BetaP[", i,"])", sep = "")

plotPost( betasample , credMass=0.95, xlab=x , ylab=py )
densidadejpeg <- paste("DensidadeBP",i ,".jpeg", sep = "")
densidade <- paste("DensidadePB",i ,".eps", sep = "")
old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()
```

```
dev.copy2eps(file=densidade ) #Guardar Densidade Posteriores
dev.copy(jpeg,file=densidadejpeg ) #Guardar Densidade Posteriores
graphics.off()
setwd(old)
getwd()

}

#Gráfico para R

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()

plotChains( "r"  , saveplots=T ,filenameroot="GraficoCadeia" )
graphics.off()

#Repor a directoria
setwd(old)
getwd()

# Verificar Densidades Posteriores com Intervalo HDI

betasample = samplesSample("r" )

x <- paste("r ", sep = "")
py <- paste("Prob(r )", sep = "")

plotPost( betasample , credMass=0.95, xlab=x , ylab=py )
densidade <- paste("Densidade_r",".eps", sep = "")
densidadejpeg <- paste("Densidade_r",".jpeg", sep = "")
old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()
dev.copy2eps(file=densidade ) #Guardar Densidade Posteriores
dev.copy(jpeg,file=densidadejpeg ) #Guardar Densidade Posteriores
graphics.off()
setwd(old)
getwd()


#Graficos de Alpha

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()

param <- paste("alpha", sep = "")
plotChains( param  , saveplots=T ,filenameroot="GraficoCadeia" )
graphics.off()

#Repor a directoria
setwd(old)
getwd()

betasample = samplesSample(param )

x <- paste("alpha", sep = "")
py <- paste("Prob(alpha)", sep = "")

plotPost( betasample , credMass=0.95, xlab=x , ylab=py )
densidadejpeg <- paste("Densidade_alpha",".jpeg", sep = "")
densidade <- paste("Densidade_alpha",".eps", sep = "")
old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()
dev.copy2eps(file=densidade ) #Guardar Densidade Posteriores
dev.copy(jpeg,file=densidadejpeg ) #Guardar Densidade Posteriores

graphics.off()
```

```
setwd(old)
getwd()

#Graficos de AlphaP

old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()

param <- paste("alphap", sep = "")
plotChains( param  , saveplots=T ,filenameroot="GraficoCadeia" )
graphics.off()

#Repor a directoria
setwd(old)
getwd()

betasample = samplesSample(param )

x <- paste("alphap", sep = "")
py <- paste("Prob(alpha)", sep = "")

plotPost( betasample , credMass=0.95, xlab=x , ylab=py )
densidadejpeg <- paste("Densidade_alphaP",".jpeg", sep = "")
densidade <- paste("Densidade_alphaP",".eps", sep = "")
old<- getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos") #Preparar aqui a pasta
getwd()
dev.copy2eps(file=densidade ) #Guardar Densidade Posteriores
dev.copy(jpeg,file=densidadejpeg ) #Guardar Densidade Posteriores

graphics.off()
setwd(old)
getwd()

#Repor a directoria
setwd(old)
getwd()
```

8.5.2 Programa Residuos ZIBN

```
getwd()
setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2") #Preparar aqui a pasta

dados_dic <- read.table("Dados_Dic.txt", header=TRUE) #tabelas com os dados sem o End ao fim do bugs

setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Resultados") #Preparar aqui a pasta

res <- read.table("ResultadoModelo.txt", header=TRUE) #tabelas com os dados sem o End ao fim do bugs

Mean.Deviance <- res [1,1]
ParamatersP1 <- res [2,1]
ParamatersP2 <- res [3,1]

Paramaters1 <- res [5:16,1]
Paramaters2 <- res [18:19,1]

r <- res [20,1]

Betas <- c(ParamatersP1 ,Paramaters1 )
Betasp <- c(ParamatersP2 ,Paramaters2 )

Betas <- c(ParamatersP1 ,Paramaters1 )
Betasp <- c(ParamatersP2 ,Paramaters2 )

# Calculo do valor dos parametros para cada individuo

#Inicializar parametros
b_intercept<- NULL
b_sexo<- NULL
b_idade<- NULL
b_h20_canal<- NULL
b_wc<- NULL
b_contacto_h2o<- NULL
b_saberschist<- NULL
b_hematuria<- NULL
b_profissao<- NULL
b_motivo<- NULL
b_provincia<- NULL
b_natura<- NULL

b<- NULL

#Inicializar parametros P
bp_intercept<- NULL
bp_sexo<- NULL
bp_idade<- NULL
bp_h20_canal<- NULL
bp_wc<- NULL
bp_contacto_h2o<- NULL
bp_saberschist<- NULL
bp_hematuria<- NULL
bp_profissao<- NULL
bp_motivo<- NULL
bp_provincia<- NULL
bp_natura<- NULL

bp<- NULL


desc_sexo<- NULL
desc_idade<- NULL
desc_h20_canal<- NULL
desc_wc<- NULL
desc_contacto_h2o<- NULL
desc_saberschist<- NULL
desc_hematuria<- NULL
desc_profissao<- NULL
```

```
desc_motivo<- NULL
desc_provincia<- NULL
desc_natura<- NULL


for(i in 1:nrow(dados_dic))
{

#Sexo
if ( dados_dic$sexo[i]== 1)
 {b='Masc.' }
else   {b='Fem.'}
desc_sexo=rbind(desc_sexo,b)

#Água Canalizada
if ( dados_dic$h2o_canalizada[i]== 0)
 {b='Não Tem'}
else   {b='Tem'}
desc_h20_canal=rbind(desc_h20_canal,b)

#Wc
if ( dados_dic$wc[i]== 1)
 {b='Fora' }
else   {b='Dentro'}
desc_wc=rbind(desc_wc,b)

#Contacto com a àgua
if ( dados_dic$contacto_h2o[i]== 2)
 {b='Lagoa' }
else if ( dados_dic$contacto_h2o[i]== 3)
    {b='Tanque'}
else   {b='Rio'}
desc_contacto_h2o=rbind(desc_contacto_h2o,b)

#Sabe da Doença
if ( dados_dic$saberschist[i]== 0)
 {b='Não Sabe' }
else   {b='Sabe'}
desc_saberschist=rbind(desc_saberschist,b)

#Mac
if ( dados_dic$hematuria[i]== 1)
 {b='Teste Positivo' }
else   {b='Teste Negativo'}
desc_hematuria=rbind(desc_hematuria,b)

#Profissão
if ( dados_dic$profissao[i]== 1)
 {b='Agricultor' }
else if ( dados_dic$profissao[i]== 3)
    {b='Doméstico'}
else if ( dados_dic$profissao[i]== 4)
 {b='Estudante'}
else if ( dados_dic$profissao[i]== 5)
 {b='Outros'}
else   {b='F. Publico'}
desc_profissao=rbind(desc_profissao,b)

#Motivo de Contacto com a água
if ( dados_dic$motivo[i]== 2)
 {b='Pescar' }
else if ( dados_dic$motivo[i]== 3)
    {b='Lav. Roupa'}
else if ( dados_dic$motivo[i]== 4)
 {b='Hig. Pessoal'}
else if ( dados_dic$motivo[i]== 5)
 {b='Nadar'}
else if ( dados_dic$motivo[i]== 6)
 {b='Outros'}
else   {b='Busc. Agua'}
```

```
desc_motivo=rbind(desc_motivo,b)

#Provincia
if ( dados_dic$provincia[i]== 2)
 {b='Luanda' }
else if ( dados_dic$provincia[i]== 3)
    {b='Bengo'}
else   {b='Kwansa Sul'}
desc_provincia=rbind(desc_provincia,b)


#Naturalidade
if ( dados_dic$natura[i]== 2)
 {b='Bengo'}
else if ( dados_dic$natura[i]== 3)
    {b='Bié, Huambo e Moxico'}
else if ( dados_dic$natura[i]== 4)
 {b='Norte'}
else   {b='Sul'}
desc_natura=rbind(desc_natura,b)
}

for(i in 1:nrow(dados_dic))
{
#Intercept
b_intercept=rbind(b_intercept,Betas[1])
bp_intercept=rbind(bp_intercept,Betasp[1])


#Sexo
if ( dados_dic$sexo[i]== 1)
 {b=Betas[2]
  bp=Betasp[2] }
else   {b=0
  bp=0}
b_sexo=rbind(b_sexo,b)
bp_sexo=rbind(bp_sexo,bp)

#Idade
 b=Betas[3]*(dados_dic$idade[i])
 b_idade=rbind(b_idade,b)

 bp=Betasp[3]*(dados_dic$idade[i])
 bp_idade=rbind(bp_idade,bp)

#Água Canalizada
#if ( dados_dic$h2o_canalizada[i]== 0)
# {b=Betas[4]
#  bp=Betasp[4]
# }
#else   {b=0
#   bp=0
# }
#b_h20_canal=rbind(b_h20_canal,b)
#bp_h20_canal=rbind(bp_h20_canal,bp)

#Wc
if ( dados_dic$wc[i]== 1)
 {b=Betas[4]
 }
else   {b=0
 }
b_wc=rbind(b_wc,b)

#Contacto com a àgua
if ( dados_dic$contacto_h2o[i]== 2)
 {b=Betas[5]
   }
else if ( dados_dic$contacto_h2o[i]== 3)
    {b=Betas[6]
```

```
 }
else   {b=0
 }
b_contacto_h2o=rbind(b_contacto_h2o,b)

#Sabe da Doença
if ( dados_dic$saberschist[i]== 0)
 {b=Betas[7]
  }
else   {b=0
  }
b_saberschist=rbind(b_saberschist,b)

#Mac
if ( dados_dic$hematuria[i]== 1)
 {b=Betas[8]
 }
else   {b=0
 }
b_hematuria=rbind(b_hematuria,b)

#Profissão
#if ( dados_dic$profissao[i]== 1)
# {b=Betas[10]
#  bp=Betasp[10]}
#else if ( dados_dic$profissao[i]== 3)
   # {b=Betas[11]
#  bp=Betasp[11]}
#else if ( dados_dic$profissao[i]== 4)
# {b=Betas[12]
#  bp=Betasp[12]}
#else if ( dados_dic$profissao[i]== 5)
# {b=Betas[13]
#  bp=Betasp[13]}
#else   {b=0
#   bp=0}
#b_profissao=rbind(b_profissao,b)
#bp_profissao=rbind(bp_profissao,bp)


#Motivo de Contacto com a água
if ( dados_dic$motivo[i]== 2)
 {b=Betas[9]
  }
else if ( dados_dic$motivo[i]== 3)
    {b=Betas[10]
  }
else if ( dados_dic$motivo[i]== 4)
 {b=Betas[11]
  }
else if ( dados_dic$motivo[i]== 5)
 {b=Betas[12]
 }
else if ( dados_dic$motivo[i]== 6)
 {b=Betas[13]
 }
else   {b=0
 }
b_motivo=rbind(b_motivo,b)

#Provincia
#if ( dados_dic$provincia[i]== 2)
# {b=Betas[19]
#  bp=Betasp[19] }
#else if ( dados_dic$provincia[i]== 3)
   # {b=Betas[20]
#  bp=Betasp[20]}
#else   {b=0
#   bp=0}
#b_provincia=rbind(b_provincia,b)
```

```
#bp_provincia=rbind(bp_provincia,bp)

#Naturalidade
#if ( dados_dic$natura[i]== 2)
# {b=Betas[21]
#  bp=Betasp[21] }
#else if ( dados_dic$natura[i]== 3)
   # {b=Betas[22]
#  bp=Betasp[22]}
#else if ( dados_dic$natura[i]== 4)
# {b=Betas[23]
#  bp=Betasp[23]}
#else   {b=0
#   bp=0}
#b_natura=rbind(b_natura,b)
#bp_natura=rbind(bp_natura,bp)
}

# Matriz com os parametros para cada individuo
Matrixparameters<- data.frame(b_intercept,b_sexo,b_idade,b_wc,b_contacto_h2o, b_saberschist,b_hematuria , b_motivo )
#write.table(Matrixparameters, "Resultados_Parametros_individuos.txt", sep=" ")

Matrixparametersp<- data.frame(bp_intercept,bp_sexo,bp_idade)
#write.table(Matrixparametersp, "Resultados_Parametros_individuos_p.txt", sep=" ")

#Soma dos parametros para cada individuo e cálculo de Lambda

lambda<- NULL  #Vector de Lambdas

tlambda<- NULL  #Vector de Lambdas
p0<- NULL #Vector de P
pstar<-NULL

for(i in 1:nrow(dados_dic))
{
lambda<-rbind(lambda,exp(sum(Matrixparameters[i,])))
tlambda<-rbind(tlambda,(sum(Matrixparameters[i,])))
p0<-rbind(p0,exp(sum(Matrixparametersp[i,])) / (1+exp(sum(Matrixparametersp[i,]))))

pstar<-rbind(pstar,r/(r+exp(sum(Matrixparameters[i,]))))
}

p100<-p0*100

#Calculo do logaritmo da probabilidade de cada individuo
prob   <- NULL
order   <- NULL

bnegative_zeros<- NULL
bnegative   <- NULL

for(i in 1:nrow(dados_dic))
{

bnegative_zeros<- NULL
if ( dados_dic$ovos10ml[i]== 0)
 { #Se Zero
 bnegative_zeros <- log (p0[i] + (1-p0[i])*pstar[i]^r )
 }
else
 { #Se maior que Zero
 bnegative_zeros <- log(1-p0[i])+ lgamma(dados_dic$ovos10ml[i]+r)-
 lgamma(dados_dic$ovos10ml[i]+1) - lgamma(r) + r*log(pstar[i]) + dados_dic$ovos10ml[i]* log(1-pstar[i])
 }
bnegative<-rbind(bnegative,bnegative_zeros)
prob<-rbind(prob,exp(bnegative_zeros))
order=rbind(order,i)
}
```

```
newframe<-cbind(dados_dic,bnegative,p0,lambda,r,prob)
p_aux <- (newframe$r/(newframe$lambda+newframe$r))^newframe$r
newframe<-cbind(dados_dic,bnegative,p0,lambda,r,prob,p_aux,r )
residual <- newframe$ovos10ml - ((1-newframe$p0) * newframe$lambda)
expected <-  ((1-newframe$p0) * newframe$lambda)
pdresidual <-(newframe$ovos10ml - ((1-newframe$p0)* newframe$lambda))
/sqrt((1-newframe$p0)*(newframe$lambda+newframe$lambda^2/newframe$r)+newframe$lambda^2*(newframe$p0^2+newframe$p0))

newframe <-cbind(dados_dic,bnegative,prob,p0,lambda,r ,residual ,expected ,pdresidual,order,desc_sexo ,desc_h20_canal,desc_wc,
desc_contacto_h2o,desc_saberschist,desc_hematuria,desc_profissao,desc_motivo,desc_provincia,desc_natura)

setwd("C:/Users/bleW/Desktop/ZINegBinamial Centrada 2/Gráficos/Residuos") #Preparar aqui a pasta
write.table(newframe, "Resultados_residuos.txt", sep=" ")

#QQplot dos residuos padronizados
jpeg('zibnNormalscoresjpeg.jpg')

qqnorm(newframe$pdresidual,  ylab="Residuos Padronizados",
 xlab="Normal Scores", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
qqline(newframe$pdresidual)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Histograma dos residuos padronizados
jpeg('zibnHistoResiduos.jpg')
max_num <- max(newframe$pdresidual)
min_num <- min(newframe$pdresidual)
mean_num <- mean(newframe$pdresidual)
hist(newframe$pdresidual, col=heat.colors(max_num), breaks=200, xlab="Residuo" ,ylab="Frequência Relativa" ,
     xlim=c(min_num ,max_num), right=F, main="Histograma de Residuos Padronizados", las=1,freq=FALSE)
title(main = "", sub = "" )
dev.off()

#Plot dos residuos para cada individuo
jpeg('zibnREsPorIndividuo.jpg')
plot(newframe$order,newframe$pdresidual , ylab="Residuos",
xlab="Indivíduo", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot das Probabilidades para cada individuo
jpeg('zibnProbPorIndividuo.jpg')
plot(newframe$order,newframe$prob, ylab="Probabilidade",
xlab="Individuo", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Valores Previstos
jpeg('zibnVsValoresPrevistos.jpg')
plot(  newframe$expected ,newframe$pdresidual, ylab="Residuos",
xlab="Valores Esperados", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1,xlim=c(0,100))
#abline(0, 0)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#QQPlot dos Valores esperados vs observados
jpeg('zibnqqploValoresPrevistosObservados.jpg')
qqplot(newframe$expected,newframe$ovos10ml,  ylab="Valores Observados",
 xlab="Valores Esperados", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
abline(0, 1)
dev.off()

#Histograma dos valores esperados
jpeg('zibnHistoEsperado.jpg')
max_num <- max(newframe$expected)
min_num <- min(newframe$expected)
mean_num <- mean(newframe$expected)
hist(newframe$expected, col=heat.colors(max_num), breaks=150, xlab="Valor Esperado" ,ylab="Frequência" ,
     xlim=c(min_num ,200), right=F, main="Histograma de Valores esperados", las=1 ,ylim=c(0,45))
```

```
title(main = "", sub = "" )
dev.off()

#Plot dos valores Previstos vs Observados
jpeg('zibnploValoresPrevistosObservados.jpg')
plot(newframe$expected,newframe$ovos10ml,  ylab="Valores Observados",
 xlab="Valores Esperados", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
abline(0, 1)
dev.off()

#Plot dos valores Previstos vs Observados por indviduo
jpeg('zibnploValoresPrevistosVSObservadosIndividuos.jpg')
plot(newframe$order,newframe$ovos10ml, ylab="Nr° de Ovos",
xlab="Individuo", main="Comparação de Valor Esperado com Observado", col="blue",type="h", pch=19 ,cex=1.5 )
 par(new=T)
plot(newframe$order,newframe$expected, ylab="", xlab="", col="red", cex=0.5, pch=19,yaxt='n')
 par(new=F)
legend("topright", legend = c("Observado"," Esperado"), col = 1:2, lty = 8)
#abline(h=mean(newframe$ovos10ml),col=5,lty=4)
dev.off()

#Plot dos residuos vs Valores Observados
jpeg('zibnVsValoresObservados.jpg')
plot(newframe$ovos10ml ,newframe$pdresidual , ylab="Residuos",
xlab="Valores Observados", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1,xlim=c(0,100))
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()


#Plot dos residuos vs Profissão
jpeg('zibnVsProfissao.jpg')
plot(newframe$profissao ,newframe$pdresidual , ylab="Residuos",
xlab="Profissão", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Idade
jpeg('zibnVsIdade.jpg')
plot(newframe$idade  , newframe$pdresidual ,ylab="Residuos",
xlab="Idade do Individuo", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Género
jpeg('zibnVsSexo.jpg')
plot(newframe$sexo ,newframe$pdresidual , ylab="Residuos",
xlab="Sexo do Individuo", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Teste Mac
jpeg('zibnVsMAC.jpg')
plot(newframe$hematuria,newframe$pdresidual , ylab="Residuos",
 xlab="Teste MAC", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Conhecimento da doença
jpeg('zibnVsConhecimentoDoença.jpg')
plot(newframe$saberschist ,newframe$pdresidual , ylab="Residuos",
xlab="Conhecimento da Doença", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Contacto com a água
jpeg('zibnVsContactoH2O.jpg')
plot(newframe$contacto ,newframe$pdresidual , ylab="Residuos",
xlab="Motivo de Contacto com Água", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
```

```
dev.off()


#Plot dos residuos vs Motivo Contacto com a água
jpeg('zibnVsMotivo.jpg')
plot(newframe$motivo ,newframe$pdresidual , ylab="Residuos",
xlab="Motivo de Contacto com Água", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Existência de WC
jpeg('zibnVsWC.jpg')
plot(newframe$wc ,newframe$pdresidual , ylab="Residuos",
 xlab="Existência de WC", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()


#Plot dos residuos vs Província de Residência
jpeg('zibnVsProvincia.jpg')
plot(newframe$provincia ,newframe$pdresidual , ylab="Residuos",
xlab="Provincia", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()

#Plot dos residuos vs Naturalidade
jpeg('zibnVsNaturalidade.jpg')
plot(newframe$natura ,newframe$pdresidual , ylab="Residuos",
 xlab="Naturalidade", main="Análise de Residuos", col=ifelse((newframe$ovos)>0, "blue", "red"), cex=1)
legend("topright", legend = c("Não Zero ","Zero"), col = 1:2, lty = 8)
dev.off()
```

8.6 Programas Auxiliares

Foram usados um conjunto de programas em adição aos criados para este trabalho. Estes estão disponibilizados no livro *Doing Bayesian Data Analysis* de John K. Kruschke (2010), capítulo 23, *Tools in the Trunk*.

8.6.1 Programa *HDIofMCMC*

Este programa permite calcular o Intervalo de Máxima Densidade de um nível de arbitrário utilizando um vector que contém uma amostra da distribuição *a posteriori*.

```
HDIofMCMC = function( sampleVec , credMass=0.95 ) {
 # Computes highest density interval from a sample of representative values,
 # estimated as shortest credible interval.
 # Arguments:
 # sampleVec
 # is a vector of representative values from a probability distribution.
 # credMass
 # is a scalar between 0 and 1, indicating the mass within the credible
 # interval that is to be estimated.
 # Value:
 # HDIlim is a vector containing the limits of the HDI
 sortedPts = sort( sampleVec )
 ciIdxInc = floor( credMass * length( sortedPts ) )
 nCIs = length( sortedPts ) - ciIdxInc
 ciWidth = rep( 0 , nCIs )
 for ( i in 1:nCIs ) {
 ciWidth[ i ] = sortedPts[ i + ciIdxInc ] - sortedPts[ i ]
 }
 HDImin = sortedPts[ which.min( ciWidth ) ]
 HDImax = sortedPts[ which.min( ciWidth ) + ciIdxInc ]
 HDIlim = c( HDImin , HDImax )
 return( HDIlim )
 }
```

8.6.2 Programa *plotPost*

O programa *plotPost* cria em R um gráfico das distribuições *a posteriori* e os respectivos Intervalos de Máxima Densidade com recurso ao programa *HDIofMCMC*.

```
 plotPost = function( paramSampleVec , credMass=0.95 , compVal=NULL ,
 HDItextPlace=0.7 , ROPE=NULL , yaxt=NULL , ylab=NULL ,
xlab=NULL , cex.lab=NULL , cex=NULL , xlim=NULL , main=NULL ,
 showMode=F , ... ) {
 # Override defaults of hist function, if not specified by user:
 # (additional arguments "..." are passed to the hist function)
 if ( is.null(xlab) ) xlab="Parameter"
 if ( is.null(cex.lab) ) cex.lab=1.5
 if ( is.null(cex) ) cex=1.4
 if ( is.null(xlim) ) xlim=range( c( compVal , paramSampleVec ) )
if ( is.null(main) ) main=""
if ( is.null(yaxt) ) yaxt="n"
if ( is.null(ylab) ) ylab=""
# Plot histogram.
par(xpd=NA)
histinfo = hist( paramSampleVec , xlab=xlab , yaxt=yaxt , ylab=ylab ,
freq=F , col="lightgrey" , border="white" ,
xlim=xlim , main=main , cex=cex , cex.lab=cex.lab ,
... )
```

```
# Display mean or mode:
if ( showMode==F ) {
meanParam = mean( paramSampleVec )
text( meanParam , .9*max(histinfo$density) ,
bquote(mean==.(signif(meanParam,3))) , adj=c(.5,0) , cex=cex )
} else {
dres = density( paramSampleVec )
modeParam = dres$x[which.max(dres$y)]
text( modeParam , .9*max(histinfo$density) ,
bquote(mode==.(signif(modeParam,3))) , adj=c(.5,0) , cex=cex )
}
# Display the comparison value.
if ( !is.null( compVal ) ) {
pcgtCompVal = round( 100 * sum( paramSampleVec > compVal )
/ length( paramSampleVec ) , 1 )
pcltCompVal = 100 - pcgtCompVal
lines( c(compVal,compVal) , c(.5*max(histinfo$density),0) ,
lty="dashed" , lwd=2 )
text( compVal , .5*max(histinfo$density) ,
bquote( .(pcltCompVal)*"% <= " *
.(signif(compVal,3)) * " < "*.(pcgtCompVal)*"%" ) ,
adj=c(pcltCompVal/100,-0.2) , cex=cex )
}
# Display the ROPE.
if ( !is.null( ROPE ) ) {
pcInROPE = ( sum( paramSampleVec > ROPE[1] & paramSampleVec < ROPE[2] )
/ length( paramSampleVec ) )
ROPEtextHt = .35*max(histinfo$density)
lines( c(ROPE[1],ROPE[1]) , c(ROPEtextHt,0) , lty="dotted" , lwd=2 )
lines( c(ROPE[2],ROPE[2]) , c(ROPEtextHt,0) , lty="dotted" , lwd=2 )
text( mean(ROPE) , ROPEtextHt ,
bquote( .(round(100*pcInROPE))*"% in ROPE" ) ,
adj=c(.5,-0.2) , cex=1 )
}

# Display the HDI.
source("HDIofMCMC.R")
HDI = HDIofMCMC( paramSampleVec , credMass )
lines( HDI , c(0,0) , lwd=4 )
text( mean(HDI) , 0 , bquote(.(100*credMass) * "% HDI" ) ,
adj=c(.5,-1.9) , cex=cex )
text( HDI[1] , 0 , bquote(.(signif(HDI[1],3))) ,
adj=c(HDItextPlace,-0.5) , cex=cex )
text( HDI[2] , 0 , bquote(.(signif(HDI[2],3))) ,
adj=c(1.0-HDItextPlace,-0.5) , cex=cex )
par(xpd=F)
return( histinfo )
}
```

8.6.3 Programa *plotChains*

Este programa produz um conjunto de gráficos associados aos testes de convergência da cadeias geradas. São gerados o histórico das cadeias, a função de autocorrelação e o gráfico da evolução da estatística de diagnóstico *potential scale reduction* $\hat{R}$ (BGR plot).

```
plotChains = function( nodename , saveplots=F , filenameroot="DeleteMe" ) {
summarytable = samplesStats(nodename)
show( summarytable )
nCompon = NROW(summarytable)
nPlotPerRow = 5
nPlotRow = ceiling(nCompon/nPlotPerRow)
nPlotCol = ceiling(nCompon/nPlotRow)
windows(3.75*nPlotCol,3.5*nPlotRow)
par( mar=c(4,4,3,1) , mgp=c(2,0.7,0) )
```

```
samplesHistory( nodename , ask=F , mfrow=c(nPlotRow,nPlotCol) ,
cex.lab=1.5 , cex.main=1.5 )
if ( saveplots ) {
dev.copy2eps( file=paste( filenameroot , toupper(nodename) ,
"history.eps" , sep="" ))

dev.copy(jpeg, file=paste( filenameroot , toupper(nodename) ,
"history.jpeg" , sep="" )) }

windows(3.75*nPlotCol,3.5*nPlotRow)
par( mar=c(4,4,3,1) , mgp=c(2,0.7,0) )
samplesAutoC( nodename , chain=1 , ask=F , mfrow=c(nPlotRow,nPlotCol) ,
cex.lab=1.5 , cex.main=1.5 )
if ( saveplots ) {
dev.copy2eps( file=paste( filenameroot , toupper(nodename) ,
"autocorr.eps" , sep="" ))
dev.copy(jpeg, file=paste( filenameroot , toupper(nodename) ,
"autocorr.jpeg" , sep="" )) }
windows(3.75*nPlotCol,3.5*nPlotRow)
par( mar=c(4,4,3,1) , mgp=c(2,0.7,0) )
samplesBgr( nodename , ask=F , mfrow=c(nPlotRow,nPlotCol) ,
cex.lab=1.5 , cex.main=1.5 )
if ( saveplots ) {
dev.copy2eps( file=paste( filenameroot , toupper(nodename) ,
"bgr.jpeg" , sep="" ))
dev.copy(jpeg, file=paste( filenameroot , toupper(nodename) ,
"bgr.jpeg" , sep="" ))}
return( summarytable )
}
```